Identifying and Assessing Interaction Knowledges, Skills, and Attributes for Future Force Soldiers

Tim Bowden
Micro Analysis and Design, Inc.

Patricia Keenan and Masayu Ramli
Human Resources Research Organization

Tonia Heffner
U.S. Army Research Institute

**United States Army Research Institute
for the Behavioral and Social Sciences**

May 2007

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**A Directorate of the Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved for distribution:

**PAUL A. GADE
Acting Technical Director**

**MICHELLE SAMS
Director**

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Stephanie T. Muraca, U.S. Army Research Institute
Jennifer Solberg, U.S. Army Research institute

NOTICES

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
May 2007	Final	Dec 2003 – February 2006

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Identifying and Assessing Interaction Knowledges, Skills, and Attributes for Future Force Soldiers	DASW01-03-C-0021
	5b. PROGRAM ELEMENT NUMBER
	665502

6. AUTHOR(S)	5c. PROJECT NUMBER
Tim Bowden (Micro Analysis and Design, Inc.); Patricia Keenan and Masayu Ramli (Human Resources Research Organization); Tonia Heffner (U.S Army Research Institute for the Behavioral and Social Sciences)	M770
	5d. TASK NUMBER
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Micro Analysis and Design, Inc 4949 Pearl East Circle Suite 300 Boulder, CO 80301	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: DAPE-ARI-RS 2511 Jefferson Davis Highway Arlington, VA 22202-3926	ARI
	11. MONITOR REPORT NUMBER
	Study Report 2007-04

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Subject Matter POC and Contracting Officer's Representative: Tonia Heffner

14. ABSTRACT *(Maximum 200 words)*:
    Developed in response to a U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) SBIR request, the Army Interpersonal Skills Assessment (AISA) battery consists of five measures designed to evaluate the interpersonal knowledge, skills and attributes (KSA) that will be required of Soldiers in the Future Force. As the Army evolves over the coming years, Soldiers will be placed in positions that require increasing interaction effectiveness. The goal of the AISA is to gauge the Soldier's aptitude to effectively manage interpersonal interactions and to identify Soldiers who may be well suited for positions where effective interpersonal KSAs may improve performance.
In Phase II of this SBIR effort, the AISA battery underwent a full development cycle including focus group reviews by senior Non-Commissioned Officers (NCOs), and pilot and field testing with the target population of first term Soldiers. The final activity in the Phase II effort was a concurrent validation where data were collected from 95 Soldiers and their supervisors in an attempt to determine the predictive ability of the AISA battery. The details of the development activities and the results of the validation effort are the subject of this report.

15. SUBJECT TERMS
SBIR Phase II Report, Future Force Soldier, Interaction/Interpersonal skills, Assessment, Selection

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	101	Ellen Kinzer Technical Publication Specialist 703/602-8047

Identifying and Assessing Interaction Knowledges, Skills, and Attributes for Future Force Soldiers

Tim Bowden
Micro Analysis and Design, Inc.

Patricia Keenan and Masayu Ramli
Human Resources Research Organization

Tonia Heffner
U.S Army Research Institute

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

ACKNOWLEDGEMENTS

IDENTIFYING AND ASSESSING INTERACTION KNOWLEDGES, SKILLS, AND ATTRIBUTES FOR FUTURE FORCE SOLDIERS: PHASE II FINAL REPORT

EXECUTIVE SUMMARY

As the Army transforms to meet future demands, Soldiers will increasingly be placed in situations that require them to demonstrate interpersonal skills, and certain jobs will evolve that may require high levels of interpersonal skills. The goal of the Army Interpersonal Skills Assessment (AISA) is to provide the Army with a method for identifying Soldiers who are likely to perform more effectively in situations that require strong interpersonal skills. This report outlines the development of the measures that comprise the AISA and discusses the validation research that was conducted to evaluate the battery's ability to predict Soldier performance.

The AISA battery contains five assessments administered in two stages. Stage One is a screening tool composed of three fully computerized measures. These measures are: (a) the Written Communication Assessment (WCA), which measures a Soldier's aptitude to effectively utilize electronic mail; (b) the Scenario Based Interpersonal Skills Evaluation (SBISE), a variant on traditional situational judgment tests (SJTs), that presents Soldiers with interpersonal situations and asks them to interpret or respond to the scene; and (c) a subset of items from the Rational Biodata Inventory (RBI) (Kilcullen, Mael, Goodwin, & Zazanis, 1999), which were used to assess Cultural Tolerance, Peer Leadership, and Diplomacy. An additional computerized measure, the Self Description Inventory (SDI) is also included for research purposes. Soldiers who "pass" Stage One move to the Stage Two assessments. Stage Two, which requires additional personnel to administer and score, consists of a semi-structured interview and two leaderless group discussions (LGD) that assess a Soldier's aptitude to relate to and lead others.

Research Requirement

The Phase I effort (Bowden, Laux, Knapp, & Keenan, 2003) identified a set of interpersonal skills and associated measures important for effective performance in the Army of the future. Having identified these assessment methods and the knowledges, skills and attributes (KSAs), the goal of the Phase II effort was to develop fully the assessment devices and conduct research aimed at validating the assessments' ability to predict Soldiers' interpersonal performance. A cyclic development process was undertaken for the assessments of the AISA battery. The development process began with collecting critical interpersonal incidents to serve as scenario material for the WCA and SBISE, and then moved to focus group reviews of the materials with senior non-commissioned officers (NCOs). After a draft set of all the measures was developed, it was pilot and field tested to provide a set of assessments that were ready for the validation effort. In the validation research, the research team collected Soldier data and supervisor ratings on 95 Soldiers.

Procedure

In the Phase II effort, each assessment underwent its own development process culminating in a single validation effort wherein all tests were administered and supervisor

ratings were collected. The first step in development was identifying relevant content that would tap the desired KSAs. With the background material in place, the first drafts of the assessments were created and reviewed by senior NCOs to refine the assessments and ensure they were appropriate for use with the target population. The refined assessments then underwent a review and revision through SME input. Finally, the AISA was subject to a validation study conducted using the test scores and supervisor ratings of 98 Soldiers.

Findings

Soldiers' "overall effectiveness" was a single overall rating provided by supervisors. Soldiers' "mean effectiveness" was an average of supervisors' ratings across 12 rating dimensions, without the overall effectiveness rating. Positive numbers indicate that Soldiers who perform well on the assessment were also rated more highly by their supervisor.

	Overall Effectiveness	Mean Effectiveness
RBI	-.15	-.06
SBISE	.15	.22*
WCA	-.08	-.21
Interview	.10	.24*
LGD (Community Center)	-.21	-.11
LGD (DC Tour)	-.19	-.10

* indicates correlation is significant at the .05 level

The results indicate that more evidence is needed before employing the AISA in a selection or assignment application. Although both the semi-structured interview and the SBISE show significant positive relationships with supervisor ratings, the lack of relationship between the other assessments and supervisor ratings must be further explored if the battery is to be used in an operational context.

Utilization and Dissemination of Findings

The results of the Phase II effort will be used to help define the potential applications of the AISA in the U.S. Army and to identify activities that would be useful in further developing the battery into a commercial quality, validated predictor of interpersonal performance applicable in both military and organizational settings.

IDENTIFYING AND ASSESSING INTERACTION KNOWLEDGES, SKILLS, AND
ATTRIBUTES FOR FUTURE FORCE SOLDIERS: PHASE II FINAL REPORT

CONTENTS

CONTENTS (continued)

List of Tables

CONTENTS (continued)

List of Figures

IDENTIFYING AND ASSESSING INTERACTION KNOWLEDGES, SKILLS, AND ATTRIBUTES FOR
FUTURE FORCE SOLDIERS: PHASE II FINAL REPORT

Chapter 1: Background and Report Organization

Background

In 2003, Micro Analysis and Design (MA&D) and the Human Resources Research Organization (HumRRO) were awarded a Phase I Small Business Innovative Research (SBIR) contract entitled "Identifying and Assessing Interaction Knowledges, Skills, and Aptitudes (KSAs) for Objective Force Soldiers." The purpose of the Phase I project was to identify the interpersonal KSAs that will be required of the Soldier of the future, and to identify or develop innovative concepts for measuring these KSAs for use in selection and assignment applications (Bowden, Laux, Keenan, & Knapp, 2003). This chapter describes the importance of this research, details the findings from Phase I of this effort, provides an overview of the assessments used in Phase II, and describes the order of the remainder of the report. Phase II assessments are described fully in the following chapters.

Importance of Identifying and Assessing Interaction KSAs

Interpersonal skills (e.g., the ability to work well in teams, to relate well to others including those from other cultures, and to act as a peer leader) are becoming increasingly more important as the roles and expectations of Soldiers expand to meet the needs of the Future Force. Working in a stabilized unit (Burlas, 2004), working on multi-national teams (Klein, Pongonis, & Klein, 2000), and working with peacekeeping and humanitarian efforts (Phillips, 2004), all require good interpersonal interactions. Ferris, Witt, and Hochwarter (2001) found that social skills are related to task performance, job dedication, and overall performance, demonstrating that, "social skill reflects interpersonal perceptiveness and the capacity to adjust one's behavior to different situational demands and to effectively influence and control the responses of others" (p. 1076). It is precisely this behavior that the researchers endeavor to explore in Phase II.

Soldiers typically approach group assignments with the expectation that their membership in that group will be short-lived. Most duty assignments are three years or less and throughout the course of a Soldier's assignment other group members are reassigned to different units in different locations. Under these conditions, if a Soldier does not work well with others, it is considered a temporary problem because the group membership will be altered. The Unit Focused Stabilization initiative (Burlas, 2004) proposes that Soldiers stay together for several years in order to reduce the disruption caused by annual reassignment. This initiative may promote greater family and community stability, and result in stronger bonds between Soldiers, even those who are having interpersonal problems. The fact that they are going to be together for years may motivate Soldiers to work out their problems (Pruitt & Rubin, 1986). Solid communication and interpersonal skills (e.g., conflict management, a strong sense of teamwork, cultural tolerance) will provide Soldiers with a framework for building and/or repairing relationships.

As part of the Global War on Terror, Soldiers assigned to duty in such areas as Afghanistan and Iraq find it necessary to perform multiple roles—warrior, peacekeeper, and humanitarian. These types of deployments require Soldiers to remain in foreign countries and to interact with both Soldiers from other nations and indigenous people for relatively long periods of time. To be most effective in their roles, it is important that deployed Soldiers understand and respect the customs and mores of the country where they are stationed (Klein et al, 2000; Phillips, 2004). Similarly, humanitarian aid and disaster relief also require many of the same skills. In these cases, Soldiers are working with people who are exhausted, frightened, and anxious. Soldiers must employ good interpersonal skills to effectively manage and assist civilians who find themselves in the midst of such crises.

Overview of Phase I Effort

Before describing the Phase II effort it is important to review the key findings of the initial research effort that occurred in Phase I of the program. The purpose of Phase I was to identify interpersonal KSAs that are relevant to Future Force Soldier performance, and to develop methods for assessing those KSAs. To accomplish these goals, four primary tasks were identified and completed.

These four tasks were:

1. Identify the interpersonal KSAs likely to be required for the Future Force Soldier
2. Research and critique measures or techniques to assess interpersonal KSAs
3. Develop a KSA-by-method measurement plan
4. Develop innovative concepts to assess the interpersonal KSAs

Identifying Interpersonal KSAs

The first step in the Phase I effort was to develop a descriptive taxonomy of Future Force Soldier interpersonal KSAs. One of the biggest challenges in developing the taxonomy was that interpersonal skills have a high degree of overlap with one another. Our task was to identify KSAs that were distinct enough to be considered independent and measurable and to make sure we captured the important facets of each (Bowden et al., 2003). The approach we took was to break complex KSAs, such as oral communication, into some of their component parts (e.g., active listening and nonverbal skills). A review of the literature covering the measurement of interpersonal skills showed that in previous research, KSAs that appeared to describe the same construct were called by different names (e.g., multi-cultural teamwork & cultural tolerance). In these cases, we adopted the name that seemed to be the most appropriate to Soldiers. Figure 1 shows how the taxonomy of interpersonal KSAs is organized. The complete list of KSAs, with definitions, is provided in Appendix A.

I.   Relating to and supporting others
        A. Ability to relate to and support peers
        B. Amicability
        C. Concern for Soldiers' quality of life

II.   Conflict management

III.   Cultural tolerance

IV.   Dependability

V.   Teamwork
        A. Team orientation
        B. Coordination
        C. Cooperativeness in problem-solving

VI.   Adaptability/Flexibility

VII.   Social Perceptiveness

VIII.   Communication ability
        A. Oral communication
        B. Active listening
        C. Nonverbal communication skills
        D. Written communication

IX.   Peer Leadership
        A. Acts as a role model
        B. Helping others
        C. Task leadership

Figure 1. Taxonomy of interpersonal KSAs

Research and Critique Measures or Techniques to Assess Interpersonal KSAs

Information in the research literature concerned with personality measurement and the experience of the project staff were used to identify and evaluate potential measurement methods. When deciding which methods to include in our assessment battery, we considered such factors as susceptibility to response distortion or faking, resources required for implementation (e.g., time, personnel), and ability to revise or develop alternate forms. The list of possible measures included commercial off-the-shelf (COTS) personality instruments, e.g., NEO Personality Inventory or 16PF Questionnaire, and measures designed for previous ARI projects, i.e., Maximizing 21st Century Noncommissioned Officer Performance (NCO21; Knapp et al, 2002) and new Predictors for Selecting and Assigning Future Army Soldiers (Select21; Knapp, Sager & Tremble, 2005),[1] as well as measures such as structured interviews and role plays that would need to be developed in Phase II. Our list of possible assessment methods is shown in Figure 2 (Bowden et al., 2003).

---

[1] The objective of the NCO21 project was to identify predictor measures to supplement the current junior NCO promotion system. The Select21 project was designed to provide personnel tests for use in selecting first term Soldiers.

Text-based
    Self-report (fixed response)
    Self-report (free response)
    Forced-choice
    Scenario-based (fixed response)

Oral interviews
    Situational, behavior description, combination or other structured
    Behavior descriptions
    Combination or other structured format
    Clinical

Simulations (computer based)
    High fidelity stimulus and response
    High fidelity stimulus and low fidelity response
    Low fidelity stimulus and high fidelity response
    Low fidelity stimulus and response

Live action
    Individual
    Group simulations
    Role play

Real life behavior
    Performance ratings
    Work product review

Figure 2. Possible assessment methods.

Develop a KSA-by-method measurement plan

Once we identified the KSAs of interest and the possible assessment methods, we created a matrix that depicted the possible methods for each KSA category. We then rated each measurement method on three criteria to determine which method would be most appropriate for measuring each KSA. The criteria used in evaluating the measurement methods were as follows:

- *Appropriateness of Method (AoM):* The degree to which that method can be used to tap the KSA. It is scored as 0 = Not Appropriate; 1 = Possibly Appropriate, 2 = Appropriate.

- *Susceptibility to Faking (F):* The degree to which the method can be easily faked. It is scored on a scale of 1 to 5 with 1 indicating highly susceptible to faking and 5 being not susceptible to faking.

- *Ease of Implementation (EoI):* The degree of difficulty associated with using the method to gauge the KSA. It is scored on scale of 1 to 5 with 1 indicating highly difficult to implement (high cost, labor intensive, etc) and 5 indicating low difficulty in implementation.

To help identify the final assessment methods, we then created a single utility index that collapsed information from the ratings listed above. It provided a score that could be used to rank order available measurement techniques.

Using the criteria listed above, we decided that one key to the success of the battery would be to reduce resources demands by using a multiple hurdle technique. This would allow a large number of candidates to be processed in the initial stage and, using the information collected in the initial phase, identify those candidates who would proceed to more resource-intensive second stage of assessment. Thus, we decided to use computer-administered measures as the first stage of measurement and to use assessments that required humans to administer and score in the second stage. The following section describes the discussion points and decisions that led to the final design of the two-stage assessment, each stage having multiple measures. Further, the following section describes in more detail the discussion topics we considered in determining the final components of the measurement method.

Develop Innovative Concepts to Assess the Interpersonal KSAs

   Effects of personality and general mental ability on performance. Following our initial discussion about possible assessment methods, we considered how best to approach development of the AISA battery. We realized it would be important to distinguish between knowing what to do in a given situation and actually applying that knowledge. In addition, knowing what to do and having the skill to use the knowledge may not always result in the expected behavior. The difference between actual performance and skill as assessed by tests is that the performance context adds additional sources of variation that are controlled in a skill assessment (Campbell, McCloy, Oppler & Sager, 1993). For example, students in a negotiation class might be able to describe the steps for interest-based bargaining on a final exam, but not be able to demonstrate negotiation skill in a real-life situation because they are distracted by the refusal of the other party to act as expected.

   Variance in skill level as assessed by a standardized measurement procedure is, in turn, a function of general mental ability (GMA), procedural knowledge relevant for the skill, and a variety of dispositional variables (e.g., personality) that are viewed as stable traits. Dispositional variables that are not stable include constructs like motivation, which is likely to vary according to the situation. However, the set of variables which make up an individual's personality are believed to be stable (Costa & McCrae, 1988, McCrae, et al., 2000 & McCrae, et al., 2002) and are important because we are interested in assessing interpersonal skills that may be constrained or enhanced by one's personality. The same would not be true for standardized assessments of technical skills. One could "know" what to do to display an interpersonal skill, but have difficulty doing it, even in a role play, because of constraints imposed by one's "personality." The model presented in Figure 3 shows our conceptual organization of the effects of GMA, knowledge, trait predispositions, and skill on performance. The solid lines are the hypothesized direct effects. The dashed lines are residual direct effects that could occur. For example, trait predispositions could have a residual effect on performance level even after accounting for their direct effect on skill level. This model served only as a method of organizing our thoughts prior to developing the instruments; it was not tested.

   Three of the instruments developed for the AISA, Written Communication Assessment (WCA), Scenario-Based Interpersonal Skills Evaluation (SBISE), and leaderless group discussions (LGDs) reflect both knowledge and skill—an understanding or knowledge of the underlying situation and skill in responding to it. The semi-structured interview asks Soldiers to recount how they have behaved in specific interpersonal situations and they are rated on their

response. The Rational Biodata Inventory (RBI) assesses trait predispositions. A Soldier may have very good knowledge of communication rules and could be expected to demonstrate that skill in the LGD. However, if he or she lacked diplomacy (one of the traits assessed in the RBI), his or her performance level might be lower than his or her actual skill.
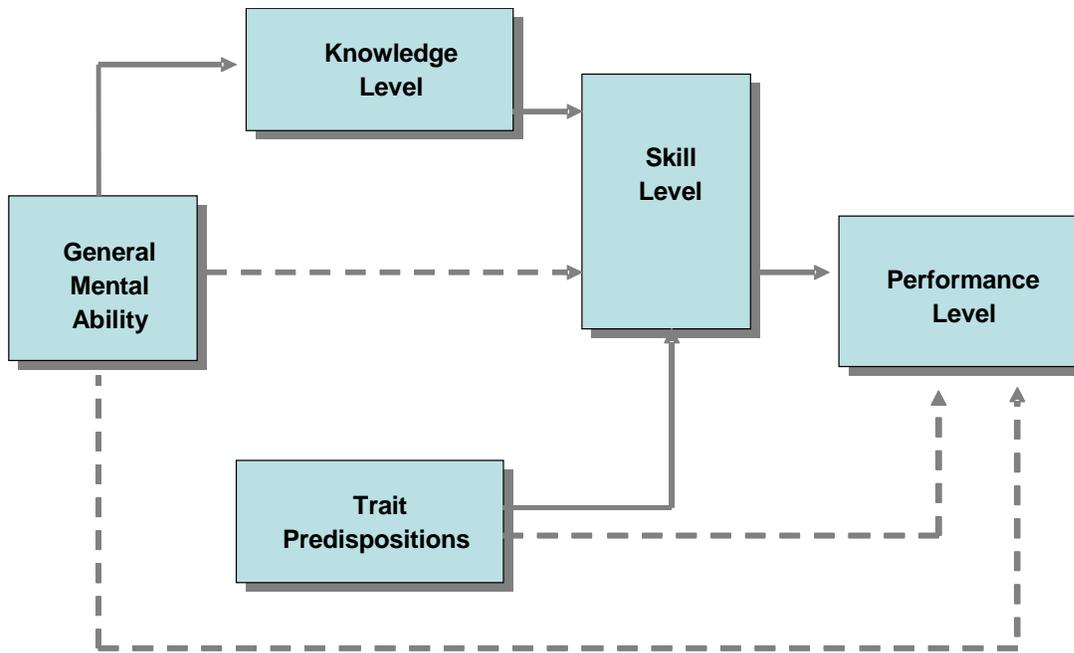
Figure 3. The hypothesized effects of general mental ability, trait predispositions, knowledge, and skill on performance.

The Army Interpersonal Skills Assessment (AISA) Battery

The Army Interpersonal Skills Assessment (AISA) battery was designed as a two-stage assessment process with three measures administered and scored via computer in Stage One, and two interactive, human-scored measures administered in Stage Two. Stage One focuses on whether an examinee *knows* what should be done when interacting with others; Stage Two assesses whether the examinee can demonstrate corresponding skill. For example, a person may know that it is inappropriate to interrupt a speaker before he or she is finished, but may still do so when interacting with others. The idea is that if an individual performs poorly in the Stage One assessments, it is not worthwhile for that individual to go to the more resource-intensive Stage Two.

Figure 4 presents a graphical view of the AISA battery. The Stage One measures include a biodata measure, a scenario-based variant of a situational judgment test (SJT), and a measure of written communication knowledge. These are computer-based and can be administered to a large number of Soldiers to determine who the best candidates for the Stage Two measures are. The Stage Two measures, which require observers and raters, include a semi-structured interview and two leaderless group discussion exercises. Each of these assessments is discussed in greater detail in the following sections.

```
┌─────────────────────────────────────────┐
│            Stage 1 Measures             │
└─────────────────────────────────────────┘

┌──────────────┐  ┌──────────────────────┐  ┌──────────────────────┐
│   Biodata    │  │ Situational Judgment │  │ Written Communications│
│  Instrument  │  │    Test/Action Exam  │  │     Assessment       │
└──────────────┘  └──────────────────────┘  └──────────────────────┘

┌─────────────────────────────────────────┐
│            Stage 2 Measures             │
└─────────────────────────────────────────┘

┌──────────────────────┐   ┌──────────────────────┐
│   Leaderless Group   │   │   Semi-Structured    │
│     Discussion       │   │      Interview       │
└──────────────────────┘   └──────────────────────┘
```
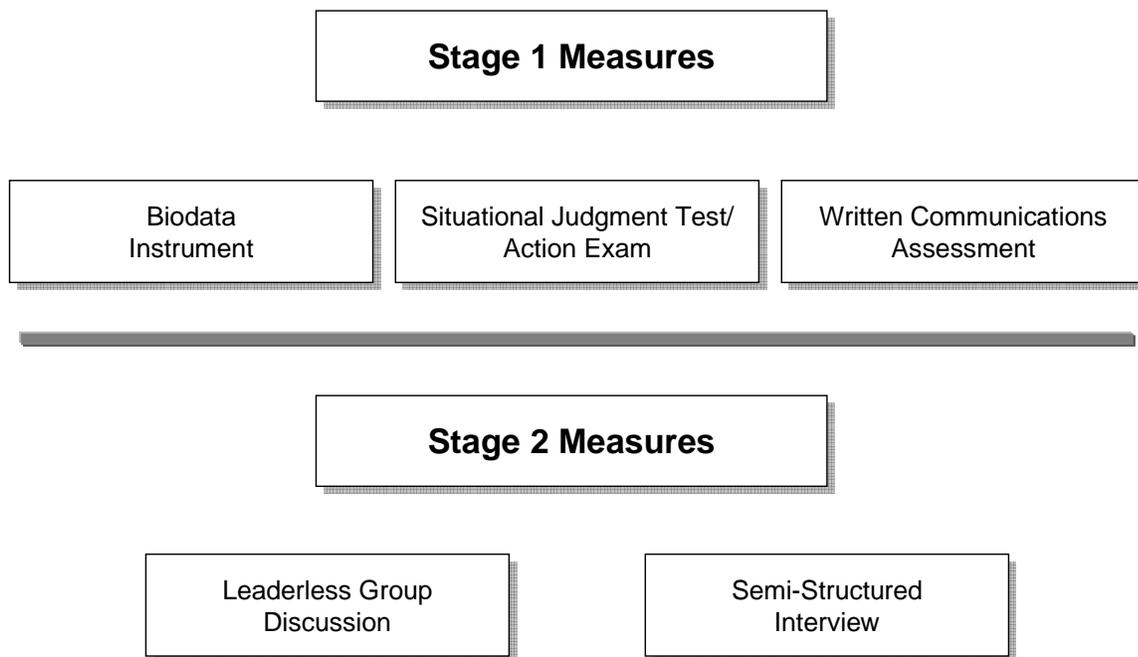
Figure 4. Graphical depiction of the stages of the AISA battery.

      Rational Biodata Inventory (RBI). Biodata tests are self-report questionnaires that use multiple-choice items to measure the test taker's prior behavior, experiences, and reactions to life events (Kilcullen, Putka, McCloy, & Van Iddekinge, 2005). Biodata items have two essential characteristics: (1) people are asked to recall and report behavior and experiences, and (2) items refer to behavior and experiences occurring in specific situations to which individuals are likely to have been exposed. Rather than develop a biodata instrument, we used a subset of the items from the RBI (Kilcullen et al, 1999). The full RBI has been shown to be a valid assessment of personality in previous research (Knapp et al., 2002, Knapp et al., 2005). It contains 16 sub-scales covering a variety of factors, many of which are not represented in our taxonomy of interpersonal skills (e.g., Cognitive Flexibility). To make the best use of available testing time, we decided to administer only the items directly related to the interpersonal KSAs targeted by the AISA. Therefore, the RBI dimensions selected for inclusion on the AISA battery are Cultural Tolerance, Peer Leadership, and Diplomacy. The modified RBI consisted of 16 multiple-choice items measuring the three dimensions.

      Scenario Based Interpersonal Skills Evaluation (SBISE). The Scenario-Based Interpersonal Skills Evaluation (SBISE) is a situational judgment test (SJT) (Motowidlo, Dunnette, & Carter, 1990) and action exam (Bigelow, 1991; Keleman, Garcia, & Lovelace, 1990; Keleman, Lovelace, & Garcia, 1991) hybrid. These are both types of assessments that our research indicated were capable of measuring several interpersonal skills. An SJT typically presents a scenario with several options for handling the situation, and then asks respondents to rate the effectiveness of each option. Like an SJT, the SBISE also presents users with a scenario and then asks them to respond to a series of questions focused on the scenario materials.

Traditionally, an action exam is used to provide candidates an opportunity to actually apply the principles learned in class (Keleman, et al., 1990; Keleman et al., 1991). It is similar to a role play in that respect, but it also allows for a discussion between the players. So, the leader may ask the role player how the dynamics of a situation might change if some aspect of the scene changes (e.g., a new person enters, one of the players responds angrily). The goal is to provide an opportunity for Soldiers to read, understand, and control a social situation (Witt & Ferris, 2003). The SBISE adopts that goal but substitutes animated characters for the role players that would be typical in an action exam.

The SBISE utilizes computer animation (see Figure 5) to present Soldiers with common interpersonal scenarios, followed by a series of questions designed to gauge each Soldier's aptitude to effectively manage interpersonal interactions. For example, Soldiers view an animation of a group of students or colleagues working together to complete an assigned project. As the scenario progresses, the video stops and the examinee is asked a variety of questions to identify a) the salient facets of the situation, b) likely outcomes given certain actions, c) factors to consider in deciding how to respond, and d) the pros and cons of possible actions. Sample questions from the SBISE include:

- Based on the actions of Jorge in the previous scenario, what word best describes his emotions?

- What things should Michelle be concerned about when deciding how to respond to Jennifer?

- What are the most likely outcomes from each of the following courses of action you could take based on the previous scenario?

Figure 5 Sample SBISE animation showing a team leader talking to a group.

Written Communication Assessment (WCA). Traditionally, assessments of writing ability include measures of writing skills such as punctuation, grammar, or vocabulary. We developed the WCA to assess the clarity of a message, both in content and tone, because we do

not consider the elements traditionally measured by tests of writing ability to be relevant to assessing interpersonal skills. With the growing popularity of electronic mail (email) as a form of communication, this format seemed a natural way to assess Soldiers' aptitude to analyze and correct written communication. Email lacks the social context cues of more traditional modes of communication, and consequently, often leads to more frequent occurrence of uninhibited behaviors (Sproull & Kiesler, 1986). As such, we expected that increased levels of interpersonal skills are required to mitigate the exhibition of such behavior.

The general format of the WCA is to present a single email or a concatenated email containing input from two individuals, then to ask several multiple-choice questions about the email. These questions might include asking Soldiers to identify (a) which of four "Subject" titles would most clearly describe the content of the message, (b) which description of the intent is most appropriate, (c) which sentences might be dropped to improve message clarity, or (d) how sentences might be reordered to improve clarity. The sentences in the emails were numbered to facilitate tasks such as reordering sentences or identifying sentences to drop. A sample item from the WCA is shown in Figure 6.

PFC Jamie Saunders is the unofficial chair of an unofficial committee that wants to buy a DVD player to use with the TV in the lobby of their enlisted quarters. He writes this note to the Sergeant in charge of the building.

SGT Griffith,

1) Everyone in our quarters wants to have a DVD player to use with the TV in the lobby. 2) Some of us who live here have taken up a collection to buy a DVD player. 3) We can get a good DVD player at the PX for the money we have collected. 4) The DVD player could be stored at the reception desk and checked out by anyone who lives in the building whenever they want to look at a DVD. 5) Some people have said they would also like to contribute to a collection of DVDs that could be stored with the DVD player and checked out, too. 6) Can we have a DVD player to use with the TV in the lobby of our quarters? 7) Let me know if this plan is OK. 8) I want to buy the DVD player as soon as possible.

1. What is the purpose of PFC Saunders' note to SGT Griffith?

2. Which sentence would make the best opening sentence for the note?

3. Which sentence(s) should be deleted to make the note more effective?

4. What is the best order of sentences to communicate most effectively?

Figure 6. Sample WCA email and follow-up questions.

Semi-structured interview. The structured interview is one of the most commonly used methods for selecting employees for hiring, training, and promotion. Structured interviews have shown to be valid in many different contexts (e.g., Campion, Pursell, & Brown, 1988; Harris, 1989; Latham, Saari, Pursell, & Campion, 1980; Pulakos, Schmitt, & Keenan, 1994), and are useful for measuring a variety of interpersonal skills, which are often difficult to gauge in other types of assessments. In addition, the interview provides an excellent opportunity to assess oral communication ability.

The semi-structured interview uses a standard protocol for conducting the interview, selecting questions from a question bank, and evaluating interviewees in several target areas. The interview is "semi-structured" in that the item pool includes multiple questions that can be asked for each KSA. The interviewers can select the questions they want to ask in a given session. Basic components of the interview include (a) a question bank, (b) rating scale forms for each KSA that include definitions and anchored rating scales for the KSA, and (c) a worksheet on which to record interviewer ratings. Two interviewers take turns asking the questions and, when the interview is completed, both raters use the rating scales to provide ratings for each of the KSA areas.

Leaderless group discussion. A leaderless group discussion (LGD) is an exercise where a small group of individuals come together to discuss a problem and reach a solution. Typically, as the discussion progresses, trained raters observe the interaction to assess participants' leadership skills (HR Guide to the Internet, 2000). The LGD is typically administered at an assessment center (where multiple simulated exercises are administered to job applicants) to measure candidates' skills and abilities.

Generally, LGDs focus on leadership characteristics (e.g., taking charge of the conversation, getting others to agree with one's position) and coming up with a final solution to a problem, which is scored by observers. The AISA LGD exercises took a different approach. They were not scored based on the quality of participants' final solutions. The exercises provided a stimulus that facilitated the assessment of participants' interpersonal skills, not just traditional leadership. Thus, several features of the LGD were designed to create situations in which participants must engage in discussion and work with other members of their group to accomplish the group's task effectively. One tactic was to give each participant different types of information: common, partially shared, and unique. Common information was provided to all participants, partially shared information was provided to more than one participant (but not all), and unique information was made available to only one participant. Our goal in doing this was to create situations in which participants had to interact with one another to uncover all the information.

We adapted existing exercises (Brockson, 1999) to develop two LGD exercises for the AISA battery. Each exercise was designed for four participants, whose behavior was scored by two trained observers. The "DC Tour" exercise directed participants to help a family structure their one-day tour of Washington, DC. The family had a list of sites they wanted to see and had received assistance from a guide at the Smithsonian. This information also provided some constraints on their schedule, for example, the panda feeding at the National Zoo, which took place at one specific time during the day.

The "Community Center" exercise asked participants to help a town plan its new community center. Participants received information about possible locations, costs for various types of facilities (e.g., weight room, internet café, hiking trail), and reactions from the townspeople. Participants were required to finalize the plans, remain within budget, and identify sources of funding to cover the down payment (25% of the cost). For both the DC Tour and Community Center exercises, a member of the group volunteered to summarize the results of the discussion to the observers.

Self Description Inventory/IPIP measure of personality. Along with the five tests that comprised the AISA battery, a sixth test was administered as part of the field test and the concurrent validation but is not a part of the final AISA battery. The Self-Description Inventory (SDI) is a 147- item instrument that measures a variety of personality-related variables. Of interest to the AISA research were the 50 items in the SDI that were selected from the International Personality Item Pool (IPIP) (International Personality Item Pool, 2001) to measure the Big Five personality factors of Agreeableness, Conscientiousness, Neuroticism, Openness to Experience, and Extraversion.[2] The Big Five is a taxonomy of personality traits—a framework for understanding which traits go together. The IPIP measure was included in the field and validation testing as a marker test of personality. A marker test is meant to provide a measure of construct validity when comparing examinees' scores on the previously validated (marker) test (in this case, the SDI) to scores on an experimental measure. If the experimental measure assesses the characteristic(s) well, then people who score high (low) on the marker test should also score high (low) on the experimental measure – that is, the scores on both measures should be positively correlated. Recall that according to the model in Figure 3, personality has a direct effect on skill and an indirect effect on performance. The results comparing the IPIP scores and AISA test scores are discussed in Chapter 9 of this report.

Overview of the AISA Development Process

In general, the AISA instrument development process (see Table 1) consisted of three steps. First, project staff developed rough drafts of the instruments. Second, these drafts were presented to NCOs, who acted as SMEs, and to Soldiers who provided input response options or suggestions for altering the exercises. Third, the instruments were pilot tested on first-term Soldiers and civilian volunteers. These development activities are outlined in the respective instrument chapters.

After the initial development, the instruments were presented to SMEs who reviewed them, offered suggestions for improving the instrument and/or provided responses that could be used as part of the assessment or for rating scales. NCOs also provided suggestions on the phrasing of instructions, background information and the overall appropriateness of the exercise for first term Soldiers. During the review process many of the assessments underwent significant changes to create instruments that were appropriate for use with first term Soldiers. Specific descriptions of development activities for each instrument are included in their individual chapters.

Pilot Tests

The SBISE and WCA were pilot tested by 33 Soldiers. The exercises were presented on laptops and used the same procedures, instructions and scenarios that were expected to be used in the final exercises. The results of these tests are discussed in detail in the individual assessment chapters that comprise the remainder of this report.

---

[2] The remaining 97 items of the SDI are composed of various interpersonal skill measures included in the battery for use in a master's thesis. The constructs measured were not of direct interest to the AISA development effort and as such the results from these items were not used in the current effort.

Table 1. Instrument Development Activities

Site	Date	# Soldiers	Activities
FT Hood	Oct 2004	20	Focus group review of SBISE, WCA, LGDs, and interview
FT Leonard Wood	Nov 2004	23	Review/revise LGDs
FT Leonard Wood	Jan 2005	36	Provide responses for SBISE and WCA
HumRRO	Apr 2005	12	Pilot test interview
FT Drum	April 2005	33	Pilot test SBISE and WCA
FT Jackson	April 2005	14	Pilot test LGDs, interview scale development
FT Riley	June 2005	35	Field test of AISA battery

The LGD exercises were pilot tested by 14 Soldiers. The exercises were videotaped also to pilot test the plan that LGDs could be conducted at any site and videotapes of the exercise could be sent to a central location for scoring.

We pilot tested the semi-structured interview with civilian volunteers. Interviewers noted questions that respondents seemed to have difficulty answering. We dropped most of these items as well as those that elicited fairly standard responses. NCOs at FT Jackson reviewed the questions and helped us develop anchors and behavioral indicators for the interview rating scales.

Field Test

Thirty-one males and four females took part in the field test. They represented 13 different military occupational specialties (MOS), which are shown in Table 2 (two Soldiers did not report their MOS).

Table 2. Composition of Field Test Sample by MOS

MOS		Frequency	Percent
13B	Cannon Crewmember	3	9.4
19D	Cavalry Scout	4	12.5
21B	Combat Engineer	9	28.1
25U	Signal Support Systems Specialist	2	6.3
35F	Special Electronic Devices Repairer	1	3.1
52D	Power-Generation Equipment Repairer	1	3.1
63B	Light Wheel Vehicle Mechanic	4	12.5
63D	Self-Propelled Field Artillery Repairer	1	3.1
63H	Track Vehicle Repairer	3	9.4
63J	Quartermaster & Chemical Equipment Repairer	1	3.1
92F	Petroleum Supply Specialists	1	3.1
92G	Food Service Operation	2	6.3
92Y	Unit Supply Specialist	1	3.1

Table 3 shows the ethnic/racial composition of the field test participants, three of whom did not report this information.

Table 3. Composition of Field Test Sample by Race/Ethnicity

Race/Ethnicity	Frequency	Valid Percent
White Non-Hispanic	22	78.6
Black	4	14.3
Asian	1	3.6
Other	1	3.6

The battery was administered in three rooms – one room for administering the computerized tests, one for LGD exercises, and one for interviews. The computer room administrator coordinated movement of Soldiers from that room to the interview and LGD. He instructed Soldiers to pause the computerized test and then report to either the interview or LGD room. He sent four Soldiers at a time to the LGD room; when they returned he sent another set of four to the LGD. Soldiers were sent one at a time to take the interview. After the field test, we made changes to the protocol, instructions, and modified various pieces of instrumentation and scoring. These changes are described in Chapter 4 for the WCA, Chapter 5 for the SBISE, Chapter 7 for the Interview and Chapter 8 for the LGD exercises.

Chapter 2: Concurrent Validation Overview

The AISA battery underwent a concurrent validation with first term Soldiers and their supervisors. The Soldiers completed all of the assessments in the AISA battery and the supervisors completed performance ratings on each Soldier. This chapter outlines the validation procedures, describes the Soldier sample and details the user experience when completing the computerized measures. Details of the AISA software, including examples of the user interface, can be found in Appendix B. Chapter 3 of this report contains detailed information on the performance rating scales supervisors used to rate their Soldiers' performance.

Testing took place in four classrooms located in the same building. Soldiers signed in and completed the computer-based portion of the test in the computer room. We administered the LGD exercises in two rooms, and used a fourth room for the Semi-Structured Interview. Two observers rated Soldiers' participation in each of the LGD exercises, and two interviewers conducted the Semi-Structured Interview.

As Soldiers took the computer-based tests, the computer room administrator randomly assigned them to take part in the other three assessments. At any one time, nine Soldiers were taking part in the LGD exercises (four at a time) or the interview. To help the administrator track completion of the various exercises, as Soldiers completed one of the other assessments (i.e., LGD or interview), they received a different colored card associated with each assessment. Soldiers were dismissed for the day when they had received all three cards and completed the computerized instruments.

Soldier Sample

A total of 99 Soldiers participated in the data collection. Eighty-five participants (86.7%) were male and 13 (13.3%) were female. Almost two-thirds (66.33%) of the sample were 11B (Infantrymen) and slightly over one-quarter (26.53%) were 88M (Motor Transport Operators), with the remainder of the sample representing a variety of MOS, as shown in Table 4.

Table 5 and Table 6 provide pay grade and race/ethnicity demographics, respectively. Across Soldiers, average time in service and average time in MOS was three years.

Table 4. Soldier Composition of Validity Sample by MOS

MOS		Frequency	Percent
11B	Infantrymen	65	66.33
15P	Aviation Operation Specialist	1	1.02
25U	Signal Support Systems Specialists	1	1.02
63B	Light Wheel Vehicle Mechanic	2	2.04
88M	Motor Transport Operators	26	26.53
92F	Petroleum Supply Specialists	2	2.04
92Y	Unit Supply Specialist	1	1.01

Note: One Soldier did not report pay grade.

Table 5. Soldier Composition of Validity Sample by Pay Grade

		Frequency	Percent
E1	Private	3	3.06
E2	Private E2	4	4.08
E3	Private First Class	12	12.24
E4	Corporal or Specialist	78	79.90

Note: Two Soldiers did not identify their pay grade.

Table 6. Soldier Composition of Validity Sample by Race/Ethnicity

Race/Ethnicity	Frequency	Percent
White, not Hispanic	54	55.1
Black	11	11.2
American Indian	2	2.0
Asian	1	1.0
Native Hawaiian/Pacific Islander	3	3.1
Hispanic/Latino	23	23.5
Multiple Selections	2	2.0

Note: Three Soldiers did not report race/ethnicity.

Stage One Assessment Overview

The three Stage One AISA assessments are computer-based measures that are administered, scored, and reported without human raters. To deliver these assessments and generate score reports, we developed a custom test development and administration software tool.

To take the computerized assessment, the user had to first launch the AISA software, and then log in by entering a unique six digit identification number that was assigned by the test administrator. Upon logging in for the first time, the user was prompted to complete the demographic identification form. Users submitted the demographic form, and then read a brief introductory text that explained the purpose and importance of the AISA measures. The user was then prompted to select the test he or she wished to complete on the assessment selection screen.

The three Stage One instruments all used a similar interface. For each item, the upper half screen displayed pertinent item-specific instructions along with any required scenario material. In the center of the screen was a text box that contained the question text, and directly under that box were response options. Along with the assessment item screen, the SBISE used an additional interface screen to display the animated scenarios. The video interface had the look and feel of commercial video players and contained standard video controls. The interface was launched when a new video segment was to be played and stayed on top of other open windows. The test taker was not allowed to close the video interface before the animation segment was completed, at which time they chose to either review the video or to return to the questions screen and respond to the relevant assessment items.

Chapter 3: Performance Rating Scales

Performance rating scales, provided by Soldiers' supervisors, were included in the validation research to provide a criterion measure for the AISA battery. Supervisors rated the full range of Soldier performance using the performance rating scales from ARI's Select21 research project (Knapp & Sager, 2005). Rating scales measuring all aspects of performance were used because it was difficult to identify specific performance measures relevant only to interpersonal skills. Even had that been possible, rating scales solely focused on interpersonal performance might have been considered irrelevant by NCOs who regularly focus on their Soldiers' full range of performance. Supervisors rated Soldiers on 12 dimensions as well as providing an Overall Effectiveness rating.

- Common Task Performance
- MOS-Specific Task Performance
- Communication Performance
- Information Management Performance
- Problem-Solving and Decision Making Performance
- Adaptation to Changes in Missions/Locations, Assignments and Situations
- Exhibits Effort and Initiative on the Job
- Demonstrates Professionalism and Personal Discipline on the Job
- Support Peers
- Exhibits Cultural Tolerance
- Demonstrates Personal And Professional Development
- Demonstrates Physical Fitness

Validation Research

Rater Training

The supervisors received a project briefing that described the rationale behind the research and emphasized the need for accurate performance ratings. The training focused on the importance of reading and using the scales provided to ensure that all raters were "on the same page" so that Soldiers would be rated against the same standard. The administrator also discussed common rating errors to make the supervisors aware of potential problems, although the emphasis was on using the scales accurately. The administrator pointed out the three performance levels (i.e., Needs Improvement, Meets Expectations, and Strength), and described how to use them to assign rating points. A sample rating scale is shown in Figure 7.

Rater Demographics

Demographic information for the 32 supervisors is presented in Tables 7, 8, and 9. Almost all supervisors rated multiple Soldiers, enabling us to collect performance data for 82 Soldiers. However, only two Soldiers were rated by more than one supervisor so interrated reliabilities could not be calculated.

C. Communication Performance		
The extent to which the Soldier speaks clearly and concisely and conveys the intended message verbally and in writing		
– Rambles or does not speak clearly	– Usually speaks clearly and concisely	– Communicates even detailed or obscure information effectively
– States idea unclearly so that the intended message is not conveyed	– Usually states ideas or information clearly so that the message is conveyed	– Conveys very detailed messages completely and accurately
– Writes documents that contain numerous, obvious errors that make the document very difficult to understand	– Writes documents that may contain punctuation or errors in grammar, but they do not interfere with understanding	– Writes documents that are virtually error-free and easy to read
Below Expectations	Meets Expectations	Exceeds Expectations
1　　　　2	3　　　4　　　5	6　　　　7

Figure 7. Sample performance rating scale.

Table 7. Composition of Supervisor Sample by Pay Grade

Pay Grade	Frequency	Percent
E4 Corporal	2	6.3
E5 Sergeant	18	56.3
E6 Staff Sergeant	9	28.1
E7 Sergeant First Class	3	9.4

Table 8. Composition of Supervisor Sample by MOS

MOS		Frequency	Percent
11	Infantryman	26	81.3
63J	Quartermaster & Chemical Equipment Repairer	1	3.1
88M	Motor Transport Operator	5	15.6

Table 9. Composite of Supervisor Sample by Race/Ethnicity

	Frequency	Percent
White	18	56.25
Black	5	15.63
Native Hawaiian/Pacific Islander	1	3.13
Other	2	6.25
Hispanic/Latino	6	18.75

Analyses of Performance Rating Data

The rating scale used a 7-point format where a rating of 1 indicated a strong need for improvement, a 4 indicated that performance met expectations, and a 7 indicated strong performance. The inter-rater reliability coefficient was .87.

Table 10 presents the descriptive statistics for each rated dimension. These results are consistent with the findings from Select21 (Knapp & Sager, 2005), in which the means for each scale were generally higher than 4.5 and standard deviations were relatively large. Cultural Tolerance, which has the highest mean rating and the lowest standard deviation of all the scales, stands out from the others (although it should be noted that the standard deviation is still greater than 1.00). Again, this is consistent with the finding from Select21, which also found low (.03 to .20) interrater reliabilities for this scale. It is possible that this scale has particularly high demand characteristics, especially for Soldiers who are expected to be tolerant of cultural differences.

Because we used the Select21 rating scales, we hoped that we would find the same three-factor structure. However, when we conducted an exploratory factor analysis of the scales, we found one predominant factor that included 10 of the performance dimensions, "Overall Performance." The second factor might be labeled "Interpersonal Skills" and includes Supports Peers and Exhibits Cultural Tolerance. The performance dimensions on the third factor are not conducive to a logical interpretation. The results of the factor analysis are presented in Table 11.

The ratings on the scales were generally highly correlated with each other, with the exception of Cultural Tolerance (see Table 12). The factor analytic and correlational results indicate that performance is predominantly unidimensional. Conversely, the results could be the results of method bias (i.e., halo error). Although the raters were trained to avoid it, halo error occurs when raters generalize an individual's performance across dimensions. So, if they generally think highly of a person, they rate him or her higher across dimensions than an unbiased observer might. If a rater does not hold the individual in high esteem, the ratings would be expected to be lower across dimensions. Accurate ratings are expected to vary across dimensions. The high correlations among scales is not an uncommon finding with rating scales; the NCO21 project had a similar result and no solid factor structure could be determined (Knapp et al., 2002). Therefore, we used two types of the performance rating scores – an average score across all of the dimensions and the scores on the individual dimensions, as appropriate. The average score differs from the overall effectiveness score provided by the supervisors.

Table 10. Descriptive Statistics for Performance Rating Scale

	N	Min	Max	Mean	SD
Common Task Performance	82	2	7	4.72	1.23
MOS-Specific Task Performance	82	2	7	4.67	1.30
Communication Performance	83	1	7	4.77	1.36
Information Management Performance	82	1	7	4.44	1.60
Problem-Solving & Decision Making Performance	81	1	7	4.37	1.67
Adaptation to Changes	80	1	7	4.66	1.47
Exhibits Effort and Initiative	81	1	7	4.31	1.56
Demonstrates Professionalism and Personal Discipline	82	1	7	4.67	1.66
Support Peers	81	1	7	5.00	1.23
Exhibits Cultural Tolerance	80	2	7	5.56	1.03
Demonstrates Personal And Professional Development	81	1	7	4.49	1.43
Demonstrates Physical Fitness	82	1	7	4.78	1.73
Overall Effectiveness	82	2	7	4.82	1.27

Table 11. Confirmatory Factor Analysis of Performance Ratings

Performance Dimension	Component		
	1	2	3
Common Task Performance	0.750	-0.132	0.198
MOS-Specific Task Performance	0.610	-0.153	0.568
Communication Performance	0.606	0.088	0.439
Information Management Performance	0.724	-0.343	-0.018
Problem Solving and Decision Making Performance	0.664	0.187	0.329
Adaptation to Changes in Missions/Locations, Assignments, and Situations	0.643	0.088	-0.185
Exhibits Effort and Initiative	0.679	0.054	-0.281
Demonstrates Professionalism and Personal Discipline on the Job	0.754	-0.187	-0.176
Supports Peers	0.551	0.563	-0.173
Exhibits Cultural Tolerance	0.277	0.831	-0.007
Demonstrates Personal and Professional Development	0.785	-0.232	-0.351
Demonstrates Physical Fitness	0.510	-0.091	-0.297
Eigenvalues	4.97	1.32	1.05

Table 12. Intercorrelations of Performance Rating Scales

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Common Task													
2 MOS-Specific Task	.46**												
3 Communication	.41**	.47**											
4 Information Mgmt	.61**	.40**	.36**										
5 Problem-Solving & Decision Making	.52**	.47**	.41**	.43**									
6 Adaptation to Changes	.52**	.22	.36**	.50**	.37**								
7 Effort and Initiative	.29**	.16	.31**	.35**	.47**	.30**							
8 Professionalism and Personal Discipline	.45**	.33**	.29**	.49**	.35**	.30**	.57**						
9 Support Peers	.33**	.20	.29**	.18	.27*	.38**	.35**	.39**					
10 Cultural Tolerance	.10	.05	.14	.03	.27*	.19	.15	.08	.44**				
11 Personal & Professional Development	.61**	.30**	.31**	.60**	.34**	.54**	.57**	.62**	.34**	.07			
12 Physical Fitness	.18	.20	.19	.32**	.34**	.35**	.27*	.37**	.29**	.01	.47**		
13 Overall Effectiveness	.71**	.54**	.59**	.71**	.69**	.68**	.66**	.71**	.58**	.29**	.77**	.57**	
14 Average Score	.72**	.58**	.52**	.69**	.72**	.67**	.60**	.69**	.54**	.29**	.77**	.47**	.851**

**: Correlation is significant at the 0.01 level (2-tailed).
 *: Correlation is significant at the 0.05 level (2-tailed).

20

Summary and Recommendations

The performance ratings have one strong performance factor and one interpersonal factor. The reason for this result is not clear, although it is similar to the results found in NCO21 (Knapp et al., 2002). In the future, the methodology should be refined to reduce the possible influence of halo error. We developed a pre-rating exercise for Select21 (Knapp et al., 2005) that involved giving raters a set of cards that had the definitions and behavioral anchors for each dimension printed on them. Before they made their ratings, the raters were asked to think of the first Soldier they were going to rate, to read the information on the cards, and to then sort cards into three piles: Strength, Adequate, and Needs Improvement. They went through this process for each Soldier they were going to rate. The Select21 field test was the first time we tried this exercise, and that was the first time that performance ratings demonstrated more than one factor. From observation, it seemed that raters were taking the task seriously and a quick scan of their ratings showed that there were differences between Soldiers. If additional validation work is to be done with the AISA battery, it would be worthwhile to have raters go through this exercise. Although the performance ratings showed low validity, making them a poor criterion, we did not have an alternative criterion measure.

Chapter 4: Written Communications Assessment

The Written Communications Assessment (WCA) aimed to measure knowledge of written communication such that participants demonstrate they can understand and interpret the tone, intent and goals of written communications sent via electronic mail. The measure was composed of a series of emails that the test taker read, and then responded to questions relating to the written materials. The WCA contained nine scenarios with an average Flesch Kincaid reading grade level of 7.6 (Flesch, 1974).

The test takers' experience in completing the WCA was similar to the procedure for completing the other computerized assessments. After logging in to the AISA system, the user chose the WCA from the selection screen, read the instructions, and then began the test. The examinee was then presented with the user interface where the email messages were displayed and the assessment items were presented (see Figure 8).

The test taker read each email message, and then read multiple-choice items related to that message. WCA items posed questions about the best arrangement of sentences to convey the desired message, the best subject line for the email, the best description of the messages' tone and intent and which, if any, of the remaining issues in the scenario were more appropriately addressed through means other than email. Test takers progressed through the nine scenarios, answering a total of 15 items, to complete the full assessment.

Instrument Development and Pilot

We began developing the WCA by identifying the types of written communications that first term Soldiers would likely encounter. Two primary categories of communications were identified: personal emails and Army related communications such as orders, post event announcements and general news postings like those seen at Army Knowledge Online. Due to the specific structure and clarity required in orders documents, they were deemed inappropriate for use in gauging a Soldier's aptitude to appropriately utilize electronic communications. Thus, the first draft of the WCA consisted of a series of email messages intended to be representative of more personal communications that Soldiers would likely encounter. The first draft of the WCA was comprised of a set of email exchanges which Soldiers read. After reading each email exchange, Soldiers responded to questions about making the messages more clear, appropriately titling the messages, and identifying the intended purpose of the messages.

Two development sessions involving detailed measure review by active duty Soldiers were conducted for the WCA. In the preliminary review, NCOs' comments on the WCA focused primarily on formatting, and they suggested changes such as adding emphasis to specific dates and times to identify message sequences, as well as ensuring that appropriate ranks were used in the communications.
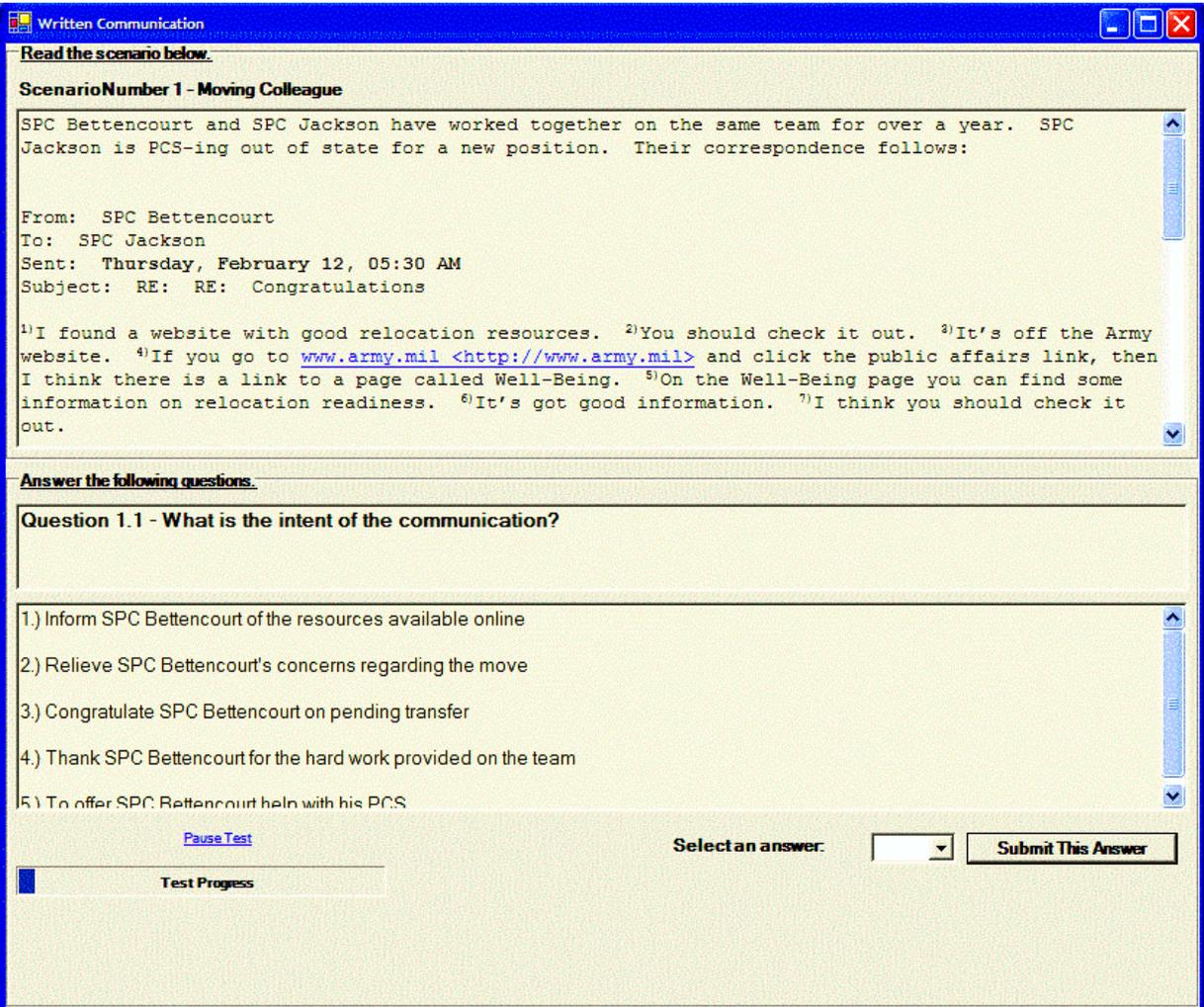
Figure 8. WCA user interface

In the follow-up session, the WCA was administered as a free response assessment to gather response options for the final multiple-choice format. In this session, Soldiers were asked to read the revised email messages, and then to compose appropriate responses to those messages. Multiple-choice response options (between eight and 10 options per item) were developed from the Soldiers' written email replies.

Twenty-seven first term Soldiers participated in a pilot test that resulted in reducing the number of response options to a range of four to eight options per item. Soldier responses were analyzed to determine how often each answer option for a particular item was selected. An initial evaluation was conducted to identify and eliminate items where 70% of more of the sample chose the same answer option. Additionally, for any item where a single answer option was not chosen by 30% or more of the sample the item was targeted for elimination or revision. Finally for remaining items answer options that were chosen by less than 10% of the sample were eliminated from the item. Overall, the number of answer options for all items was reduced to between four and eight. Additionally, one item was substantially revised by combining some of

its potential answer options, and by clarifying the questions. No items were deleted based on the aforementioned criteria.

Scoring the WCA

The scoring key for the WCA was developed using subject matter experts from the contractor team. The contractor team acknowledges that identifying subject matter experts in regard to their written communication ability can be a difficult task. As such, the six members of the SME panel were chosen because they were individuals considered by colleagues to be effective users of written communications. These SMEs were considered by colleagues to use email appropriately, to compose messages that are characteristically clear and concise and to have a good understanding of how to clearly convey the intent of their communication. Additionally, each of the SMEs had a background in the area of personality assessment so they understood the general approach and theory underlying the assessment. This expertise enabled them to make educated judgments about the assessment items in the context of interpersonal skills assessment. Each of the six participants was asked to complete the WCA, yielding six output files for analysis. The correlations between each rater can be found in Table 17. Once the SMEs completed the WCA individually, they took part in a conference call to review the assessment and obtain a consensus answer for each item. During the conference call, the SME group evaluated each assessment item and, for multiple-choice items, determined which of the response options was the correct response. The SME-designated correct option was assigned a value of one point when scoring the WCA output files for an individual Soldier.

Table 13 Inter-rater Correlations for WCA Keying

Rater	1	2	3	4	5
2	.43*				
3	.45*	.44*			
4	.41*	.56**	.47**		
5	.37*	.70**	.21	.59**	
6	.44*	.35	.56**	.42*	.22

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed)

Instrument Validation

Results

Correlation coefficients were calculated to examine the relationship between the mean of all supervisor rated dimensions ($M = 4.7$, $SD = .92$) and overall scores on the WCA ($M = 4.30$, $SD = 2.28$). Analyses revealed no relationship between performance on the WCA and supervisor ratings of overall effectiveness ($r = -.21$, *n.s.*). Additionally, no relationship was found between overall WCA score and supervisor ratings of Communications performance ($r = .02$, *n.s.*). However, it is important to note that the supervisor rating of Communication performance covered a wider performance domain that included both written and oral communications aspects.

The lack of significant correlations between supervisor ratings of both overall performance and Communications performance suggests the need for further investigation into the validity of the WCA. Several factors should be considered in interpreting the results of the analyses. First, the criterion measure, in particular the ratings of Communications performance, may not be valid because supervisors of first term Soldiers, like those in the validation sample, likely have little opportunity to observe the written communications of their subordinates. Second, very little work has been done in attempting to measure individual differences in terms of email ability. The goal of the WCA was to assess a Soldier's aptitude to correctly interpret the tone and intent of email. However, there is little literature available that describes interpreting tone and intent of email, and as such it is possible that the WCA failed to represent characteristics that are salient in evaluating tone and content of an email message. Third, as mentioned previously in this section, the criterion measure used for these correlations was composed of a larger performance domain than that measured by the WCA. This mismatch between criterion and predictor measures may mean that although WCA failed to measure the same variables as the criterion measure it may still measure the desired traits, but that no conclusion in this regard can be made based on available data. Finally, the construct the WCA intended to measure written communication skill by using a format that would be familiar to Soldiers, such as email. However, it is possible that the measurement technique did not fully capture the underlying construct.

In addition to reviewing the relationships described above, the WCA was evaluated for potential differential impact on subgroups of interest in the validation sample. Table 14 to Table 16 below present the descriptive statistics for the subscales and overall scores on the WCA for relevant subgroups. Independent samples t-tests were conducted to compare the means between the groups identified in the tables, with no significant mean score differences found.

Summary and Recommendations

The WCA attempted to measure a Soldier's aptitude to effectively interpret the interpersonal aspects of electronic mail. It targeted aspects of email including the understanding of message tone, intent, and how to improve the conveyance of these message aspects. While a great deal of research has been devoted to determining the effects of electronic communication's removal of social context cues (Sproull & Kiesler, 1991; McCormick & McCormick, 1992), little has been done in the way of measuring individual differences in how the removal of those cues impacts effective use of written communications. Future efforts should be focused on more specific identification of the aspects of electronic and other written communication that are relevant for identifying and interpreting the tone and intent of a communication. By pinpointing the salient aspects of electronic communications, the WCA could be more effectively targeted at assessing a Soldier's aptitude to identify and employ important message aspects that aide in tone and intent identification.

Additionally, while preliminary Soldier reviews supported the emails utilized in the WCA, some participants reported that the realism of the stimuli was suspect because the language and length were not typical of the communication of first term Soldiers. It is possible that the WCA's target audience of first term Soldiers do very little electronic communication in the job setting and such behaviors are restricted to personal topics. The confluence of Army communication styles, which emphasize succinctness and clarity, and the primary use of

electronic communication for personal messages may produce email that is less formal and more direct than the stimuli created by WCA developers. As such, the stimuli may have been sufficiently unfamiliar to Soldiers to make it difficult for them to identify the interpersonal factors of interest. Also, stimulus unfamiliarity may mean that high performance on the current WCA may not necessarily be related to effective written communication performance in an Army context, as evidenced by the lack of correlation between supervisor ratings of Communications performance and performance on the WCA.

Table 14. Summary Statistics for WCA Score by MOS Group

	All					11B			All Other MOS		
	N	Min	Max	Mean	SD	N	Mean	SD	N	Mean	SD
WCA Score	92	1.5	12.00	5.61	2.51	61	5.52	2.44	31	5.79	2.44

Note. Scores on the WCA are on a 15 point scale

Table 15. Summary Statistics for WCA Score by Pay Grade

	All					E4			E1-3		
	N	Min	Max	Mean	SD	N	Mean	SD	N	Mean	SD
WCA Score	91	1.5	12.00	5.61	2.51	74	5.51	2.40	17	6.24	2.93

Table 16. Summary Statistics for WCA Scores by Race/Ethnicity

		N	Mean	SD	$F_{(2,79)}$
WCA Score	White	54	5.51	2.56	.30
	Black	11	5.26	2.70	
	Hispanic/Latino	23	5.61	2.42	

Note. Comparisons of individual pairs of sub-group means show no significant differences.

Chapter 5: Scenario-Based Interpersonal Skills Evaluation (SBISE)

The Scenario Based Interpersonal Skills Evaluation is one of the three computerized tests administered as part of the AISA battery. As mentioned previously in this report, the SBISE is an SJT-variant that uses computer animations to present interpersonal scenarios. The SBISE asks test takers to interpret facial expressions, body language and other visual cues to identify the emotional state of characters depicted in the animation.

The constructs measured by the SBISE are based on the measurement plan developed under the Phase I effort. The five constructs allocated to the SBISE are: Cultural Tolerance, Social Perceptiveness, Concern for Soldier Quality of Life, Conflict Management and Peer Leadership. Each scenario in the SBISE targets between one and three of these constructs. In Chapter 1 we described the Phase I activity that was used to determine which scales would be addressed by each specific measure. Using the Appropriateness of Measure calculations we determined that for these scales the SJT type approach would be an appropriate measurement method. Additionally, for Social Perceptiveness and cultural tolerance, we believed that the use of the video stimulus would allow for asking questions that would tap specific aspects of these dimensions that would be difficult to gauge using other methods like the interview or leaderless group discussion. For example, we define Social Perceptiveness as the degree to which an individual is able to monitor own and other's emotions, discriminate among them, and use the information to guide one's thinking and actions, allowing one to work cooperatively with others.

To access the SBISE, the user logged in and selected "Start Scenario Based Test" from the assessment selection screen. The user then read the instructions and proceeded to open video player screen (see Figure 9). Users then viewed the first animated scenario of the SBISE. Each scenario consisted of two to four segments. Users had to watch a full video segment before they could answer any of the test items associated with the particular segment. After the user answered all items associated with a particular video segment, the software proceeded to the next segment within the scenario until all segments and their corresponding assessment items were completed. When the user completed the items for a scenario, the software proceeded with the first segment of the next scenario and the process continued until all scenarios and their associated items were completed.

Figure 9. Sample SBISE video player interface.

After viewing the animation segment, the users were presented with a series of assessment items that asked them to make decisions based on relevant interpersonal factors depicted in the scenario (see Figure 10). The SBISE contained two primary item-types: multiple-choice and rate-type items. Multiple-choice items were comprised of the item stem with between four and six response options. For multiple-choice items, the test taker was asked to select the best option of those presented in the given animated situation. Rate-type items presented the test taker with several possible responses to the animated scenario, and the test taker then rated the effectiveness of each potential response on a five-point Likert scale.
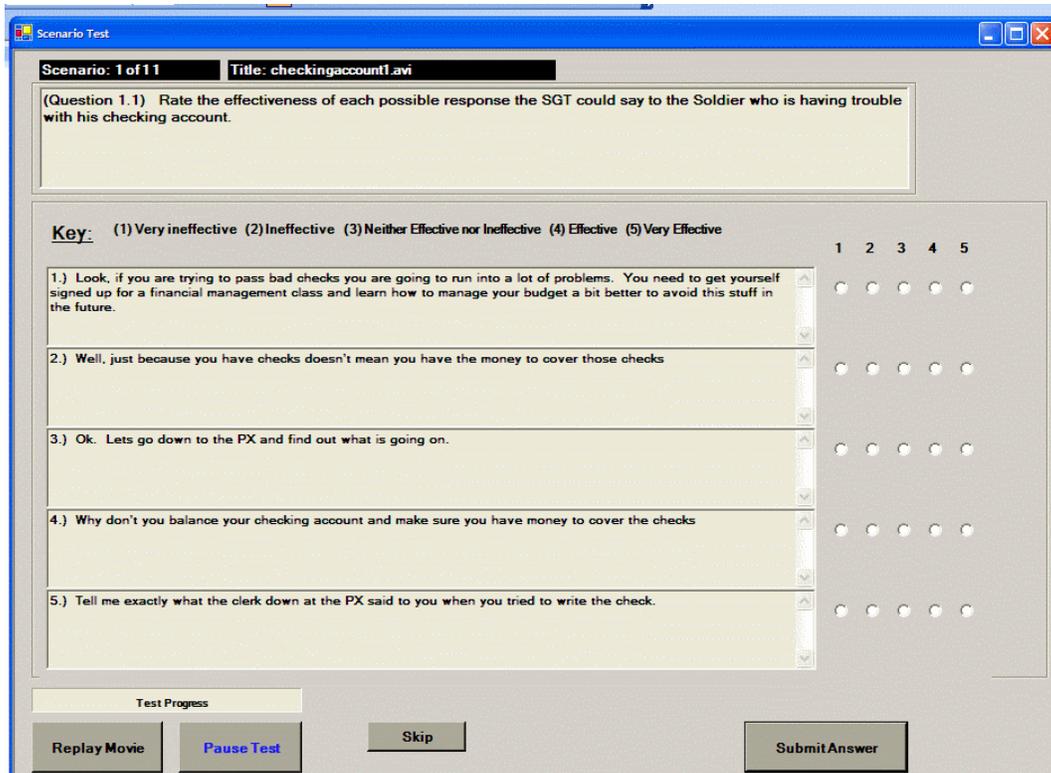
Figure 10. SBISE items interface.

Instrument Development and Pilot

The SBISE development process began by identifying potential interpersonal events that could be used as the scenario content. Several sources were reviewed for descriptive incidents that would prompt interactions between first term Soldiers. These sources included web content from Army related sites such as companycommand.com (http://companycommand.arm.mil), Army Knowledge Online (AKO - http://www.army.mil/ako/) and the Center for Army Lessons Learned (CALL - http://call.army.mil/), critical incidents gathered in previous Army projects, and discussions with former Army personnel. Using ideas collected from these sources, a set of story boards and dialogues was created that would be used to generate the full animated scenarios. Scenarios were scripted, and first draft versions of the scenarios were created. These first drafts took the form of roughly animated comic strip-like displays that depicted the actions for the scenario and contained the audio tracks for the final animations. Figure 11 shows the first-stage animatics. The animatic presentations were used in each data collection with the exception of the validation when the final animations were completed.
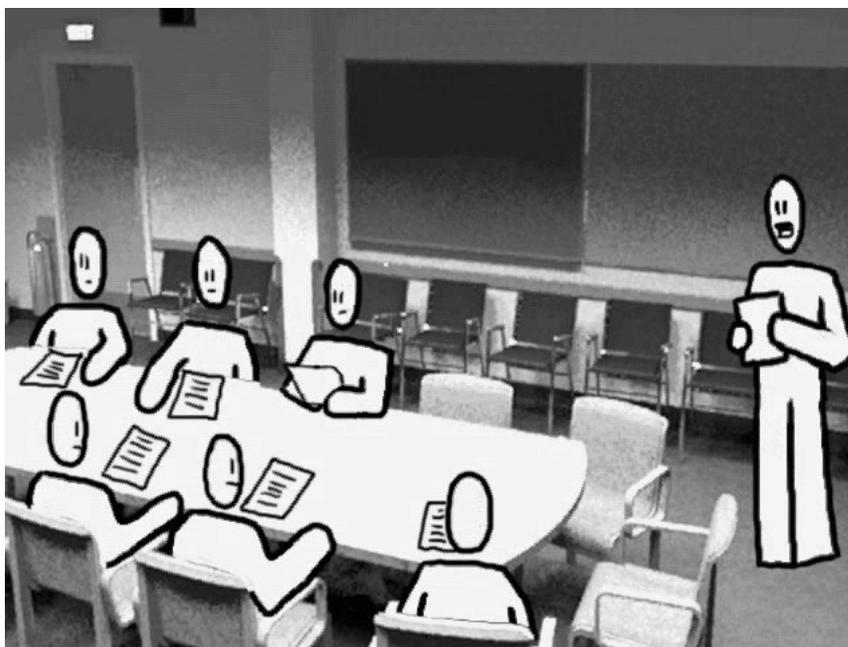
Figure 11. Sample SBISE animatic showing a team leader talking to group.

Two development sessions were conducted for the SBISE. First, NCOs were asked to review text descriptions of the scenarios and provide feedback on appropriateness for first term Soldiers. Second, 36 first term Soldiers watched animatics of several interpersonal scenarios, viewed the test items, and provided a free response that represented the best course of action. The responses generated by these Soldiers were used to create answer options for the items.

The goal of the pilot testing was to reduce or revise the SBISE answer options. Thirty-one first term Soldiers completed the measures. Rate-type items were presented with up to 10 response options, and multiple-choice items were presented with between four and eight answer options. To refine the items, researchers determined how many Soldiers selected each answer option for a particular item. A set of predetermined criteria were then applied to each item to reduce the number of responses to the goal of five for rate items and between four and six for multiple-choice items. Response options that were rated the same by more than 70% or fewer than 30% of the Soldiers were the first target for elimination. So, if more than 70% or less than 30% of Soldiers responded that an option was rated as a 3 (on a five point scale) then the item was targeted for elimination. For multiple-choice type items, response options that were chosen by less than 10% of the sample were eliminated from the item. Second, items where 70% or more of the sample chose the same response options were removed. If no option was chosen by 30% or more of the sample, the item was targeted for either elimination or revision.

Based on these predetermined criteria, the total number of assessment items was reduced to 41. Twenty-seven of the SBISE items were multiple-choice questions with a maximum point value of 1 point for correctly choosing the best available answer. The remaining 14 items were rate-type items where test takers rated up to 5 individual responses to the scenario. For each of the ratings provided by the test taker between 0 and 1 point was awarded based on the distance between the test takers rating and the rating assigned the particular item by the SME scoring

panel, as described in the following section of this report. The combination of items (rate and multiple-choice) yielded a total of 91 opportunities for the test taker to earn points on the SBISE.

Final SBISE

The final SBISE consists of 10 animated scenarios. The scenarios ranged from 1 minute 30 seconds to 3 minutes in length with 91 opportunities for Soldier to respond.

SBISE Scoring

The scoring key for the SBISE was developed using subject matter experts from the contractor team. The SME panel used in the SBISE keying activity was the same panel of experts utilized for WCA development. The six SME participants completed the SBISE which yielded six output files for analysis. The correlations between each rater can be found in Table 17. The numbers represent the degree of relationship between answers chosen for each individual item by each SME. Once the SMEs completed the SBISE individually, they reviewed the assessment and obtain a consensus answer for each item. The SME group evaluated each assessment item and for multiple-choice items determined which of the response options was correct. The SME-designated correct option was assigned a value of one point when scoring the SBISE output files for an individual Soldier.

Table 17. Interrater Correlations for SBISE Keying

Rater	1	2	3	4	5
2	.30**				
3	.56**	.30**			
4	.43**	.26*	.40**		
5	.36**	.21	.41**	.34*	
6	.49**	.28**	.50**	.45**	.43**

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed)

For rate items, the SMEs arrived at a consensus rating for each response option provided. That is, SME agreement was obtained on the effectiveness of each item and this agreed upon rating was used as the correct rating for the particular item. Using this correct rating, two scoring outputs are provided in the SBISE. First, for an overall test score, test takers are awarded one point for selecting the same rating as that determined by the SMEs. For selections that differ from the SME, Soldiers are given fractions of points based on how far their selection is from that of the SMEs. So, the point value assigned to a Soldier for a rate type item is defined as:

$$\frac{SME\ Answer - \left| User\ Answer - SME\ Answer \right|}{\#\ Possible\ Points}$$

The point value obtained from this equation is then added to the points earned on multiple-choice items to calculate the assessment score. The SME-generated key and the scoring approach described above was used to derive participant scores on the SBISE for the validation effort that is the subject of the remainder of this chapter.

Results

The SBISE contained a number of target interpersonal KSAs which were evaluated for their relationship both to supervisor ratings of overall performance and other AISA measures of the same KSAs. The scales calculated for the SBISE were Cultural Tolerance (4 items), Social Perceptiveness (9 items), Concern for Soldier Quality of Life (7 items), Conflict Management (37 items) and Peer Leadership (18 items). Internal consistency estimates for each scale were calculated and are presented along the diagonal in Table 18, along with the correlations between scale scores.

Table 18. SBISE Scale Reliabilities and Correlations

	Mean	SD	1	2	3	4	5	6
1-Concern for Quality of Life	4.52	.69	.19					
2-Conflict Management	24.55	3.06	.38**	.50				
3-Cultural Tolerance	1.67	.56	-.14	-.14	.58			
4-Social Perceptiveness	2.57	1.49	.02	.13	-.04	.20		
5-Peer Leadership	10.31	1.75	.17	.35**	-.08	.11	.45	
6-SBISE Total Score	43.00	4.98	.47**	.85**	.01	.42**	.65**	.48

** Correlation is significant at the 0.01 level (2-tailed).
n = 95 except n = 91 for Cultural Tolerance.
Numbers along diagonal are α.

The reliability analysis indicates that alpha internal estimates of reliability values fall short of desired criteria. One potential explanation for the decreased reliability of the Concern for Soldier Quality of Life, Cultural Tolerance, and Social Perceptiveness scales is the small number of items used in the scale. The small number of items increases the chance that random variability and error variance were being captured in the scale scores and reduces the likelihood of capturing true score performance on this scale. Another possible explanation is that, despite our best attempts, the constructs overlap because they are so similar. Even though the items were developed to represent the constructs, the realistic responses generated by SMEs may not be quite so clean.

The relationship between SBISE total score and supervisor overall effectiveness ratings was not significant ($r = .15$, *n.s.*), however SBISE total score and the mean ratings from all supervisor rated dimensions were positively correlated ($r = .22$, $p = .05$). Correlations between the supervisor ratings of Tolerance and SBISE Cultural Tolerance scale scores were not significant ($r = .01$, *n.s.*).

The significant positive relationship between supervisor mean ratings of effectiveness and overall SBISE score is a promising result, but must be interpreted with caution because of the low reliabilities. The relationship suggests that Soldiers who perform better on the SBISE are also likely to be rated as more effective by their supervisors. While the correlation is relatively low (r = .22) the result still suggests that an overall score from the SBISE may be effective for predicting Soldier performance and interpersonal skill.

In addition to reviewing the relationships described above we evaluated the SBISE for potential differential impact on subgroups of interest in the validation sample. Table 19 and 20 present the descriptive statistics for the subscales and overall scores on the SBISE for relevant subgroups. Independent samples t-tests were run to compare the means between the groups identified in the tables, with no significant mean score differences found.

Table 19. SBISE Scale Summary Statistics by MOS

	All					11B			Other MOS		
	N	Min	Max	Mean	SD	N	Mean	SD	N	Mean	SD
Concern for Quality of Life	95	1.80	6.00	4.52	.69	64	4.45	.71	31	4.67	.61
Conflict Management	95	4.00	29.60	24.55	3.06	64	24.51	3.38	31	24.63	2.31
Cultural Tolerance	91	.60	3.00	1.67	.56	60	1.68	.55	31	1.63	.57
Social Perceptiveness	95	.00	6.00	2.57	1.49	64	2.47	1.36	31	2.79	1.73
Task Leadership	95	6.60	15.20	10.31	1.75	64	10.37	1.81	31	10.19	1.64
Total Points	95	13.00	55.10	43.88	4.98	64	43.69	5.42	31	44.27	3.96

Note. No significant differences were found between sub-group means. Ranges for each of the SBISE scales is as follows: Concern for Quality of life: 0 – 15; Conflict Management: 0 – 35; Cultural Tolerance: 0 – 3; Social Perceptiveness: 0 – 11; Task Leadership: 0 – 19; Total Points: 0 - 83.

Table 20. Summary Statistics for SBISE Scales by Rank

	All					E4			E1-3		
	N	Min	Max	Mean	SD	N	Mean	SD	N	Mean	SD
Concern for Quality of Life	94	1.80	6.00	4.52	.69	76	4.57	3.59	18	4.06	.91
Conflict Management	94	4.00	29.60	24.55	3.06	76	24.66	2.19	18	24.09	5.51
Cultural Tolerance	94	0.60	3.00	1.67	.56	73	1.64	.51	17	1.81	.73
Social Perceptiveness	90	.00	6.00	2.57	1.49	76	2.44	1.41	18	3.23	1.58
Task Leadership	94	6.60	15.20	10.31	1.75	76	10.35	1.75	18	10.18	1.87
Total Points	94	13.00	55.10	43.88	4.98	76	43.89	3.59	18	44.06	4.99

Note. No significant differences were found between sub-group means.

Table 21. SBISE Score Summary Statistics for Race/Ethnic Subgroups

		N	Mean	SD	$F_{(4,89)}$
Concern for Quality of Life	White	54	4.45	.76	.47
	Black	11	4.60	.70	
	Hispanic/Latino	23	4.68	.50	
Conflict Management	White	54	24.65	3.64	.57
	Black	11	24.47	1.91	
	Hispanic/Latino	23	24.88	2.45	
Cultural Tolerance	White	54	1.66	.47	1.41
	Black	11	1.53	.72	
	Hispanic/Latino	23	1.79	.68	
Social Perceptiveness	White	54	2.66	1.57	.50
	Black	11	2.38	1.48	
	Hispanic/Latino	23	2.58	1.17	
Task Leadership	White	54	10.43	1.77	1.24
	Black	11	9.78	1.43	
	Hispanic/Latino	23	10.24	1.64	
Total Points	White	54	44.08	5.88	1.02
	Black	11	42.93	4.20	
	Hispanic/Latino	23	44.51	3.66	

Note. Comparisons of individual pairs of sub-group means show no significant differences.

Summary and Recommendations

The results of the SBISE validation against supervisor ratings of effectiveness are somewhat ambiguous. While significant positive correlations were found between mean effectiveness ratings and overall SBISE performance, a similar relationship between the supervisors' overall effectiveness rating and SBISE performance was not found. This could, in part, be due to the mid-level reliability estimates found among SBISE scales. Additionally, there were sample characteristics which may have impacted the data collected. For example, nearly all of the Soldiers utilized in the validation sample had recently returned from combat missions so it was possible that their interpersonal skills were altered due to the environment. That is, Soldiers may only exhibit a specific set of interpersonal behaviors in a hostile combat environment and these behaviors are likely to be governed by more strict rules and procedures than would be observed in more typical environments. As such, both the Soldier responses and their supervisors' ratings may have been influenced by the unique interpersonal conditions of a combat mission.

Another factor that could potentially impact SBISE performance has to do with the use of cartoon-like animations to depict the interpersonal scenarios. According to Arvey, Strickland, Drauden and Martin (1990) motivational factors can have an impact on overall test performance. As such, it is possible that if participants did not find the animations to be realistic or valid representations of common Army scenarios that the resulting lack of motivation to perform may have impacted scores on the SBISE. Previous research (Macan, Avedon, Paese, & Smith, 1994; Schmidt, Greenthal, Hunger, Berner & Seaton, 1977; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993) has shown that selection assessments involving simulations are viewed more favorably than paper-and-pencil assessments. Additionally, Chan and Schmitt (1997) found that

a video-based SJT had a higher perceived face validity than an equivalent paper-and-pencil measure. These factors seem to indicate that the impacts of validity perceptions on participant motivation are likely to be more favorable than when using a paper-and-pencil measure. The closest analog to the SBISE in the literature reviewed was the assessment used by Chan and Schmitt (1997) which uses live action video footage for the assessment, still does not provide adequate assurance that the computer animations used in the SBISE are perceived as face valid by participants. During data collection activities general comments from participants indicated that interest in and motivation to complete the animated scenarios was reasonably high. Additionally, logic suggests that due to the prevalence of computer animations used in motion picture productions and video gaming, participants should have had prior exposure to this type of technology. However, no systematic measurement of participant perceptions of the animations was completed in the current effort. Currently, more research is needed to explore participant perceptions of the animated scenarios and the impacts these perceptions may have on motivation and other performance moderating constructs.

While the moderate relationships between SBISE scores and supervisor ratings, in conjunction with the low reliability estimates limits the applicability of the measure for selection and assignment, there is potential for using the instrument as a means for identifying developmental areas for Soldiers. The score variance across measured constructs suggests that differences are being identified, and this coupled with the content validity evidence based on NCO approval of the scenarios and items suggests that the measure may tap the constructs desired. Given these factors, there is potential that Soldiers who complete the assessment could receive valuable feedback on KSAs where there is potential for increased development.

Future work with the SBISE should focus on gathering additional data to support the relationship between SBISE score and supervisor ratings of effectiveness. Future efforts should be undertaken to validate the SBISE on a larger, more diverse population of Soldiers. Additionally, the weakest KSA measure that is included in the SBISE is the measure of Cultural Tolerance. The current effort reveals no relationship between SBISE measures of cultural tolerance and any criteria or predictor measures. Additional investigation into the use of the SBISE as a measure of Cultural Tolerance is warranted to determine if the use of a scenario based test can accurately predict a Soldier's aptitude to interact effectively with those of other cultural backgrounds. Finally, continued research with the SBISE should examine the face validity perceptions participants have of the animated scenarios, compare those perceptions to an equivalent assessment that employs text or live action video for scenario presentation and also evaluate the impact of validity perceptions on motivation and test performance.

Chapter 6: Rational Biodata Inventory

As described in the AISA Overview portion of this report, a subset of items from the Rational Biodata Inventory (RBI) were used as part of the AISA battery. Biodata tests are self-report questionnaires that use multiple-choice items to measure the test taker's prior behavior, experiences, and reactions to life events (Kilcullen et al., 2005). Biodata items have two essential characteristics: (1) people are asked to recall and report behavior and experiences, and (2) items refer to behavior and experiences occurring in specific situations to which individuals are likely to have been exposed.

In all, three subscales of the RBI are utilized for the AISA battery. These scales are: Cultural Tolerance, Peer Leadership and Diplomacy. These three scales were selected because they most closely relate to the KSAs of interest in the AISA battery. Development activities conducted for the RBI were minimal because the assessment was previously developed and validated (Kilcullen et al., 2005) under other Army programs. The RBI version used as part of the AISA contains a total of 16 items measuring the three KSAs of Diplomacy (5 items), Peer Leadership (6 items) and Cultural Tolerance (5 items).

Instrument Development and Pilot

The primary development activity associated with the RBI was a review of the existing items and the constructs they were intended to measure to ensure that similar operational definitions were used for the RBI and the current effort. The review indicated that the RBI scales listed previously contained items measuring similar aspects of the target traits of interest in the AISA. As such, the items were adapted for computer administration as part of the AISA software. The RBI was not implemented as part of the AISA battery until the field testing described in Chapter 1 of this report, and as such no formal pilot of the revised instrument was conducted.

Instrument Validation

Results

Table 22 contains the descriptive statistics for total RBI score and the three subscales from the used in the AISA battery. Additionally, Table 23 presents the inter-correlations and internal consistency estimates for RBI total score and the three RBI subscales.

Table 22. Descriptive Statistics for RBI

	N	Minimum	Maximum	Mean	SD
Total Score	92	34.00	75.00	56.37	8.29
Diplomacy	92	9.00	25.00	18.13	3.77
Peer Leadership	92	9.00	30.00	19.89	3.57
Cultural Tolerance	92	11.00	25.00	18.55	3.97

Table 23. Scale Reliabilities and Inter-correlations for the RBI

	1	2	3	4
1 - Total Score	.77			
2 - Diplomacy	.81**	.69		
3 - Peer Leadership	.70**	.46**	.61	
4 - Cultural Tolerance	.71**	.36**	.16	.71

Note: Numbers on the diagonal represent internal consistency estimates (α).
  $n = 91$; ** indicates correlation is significant at .01 level.

The reliability analysis indicates that for all scales (Diplomacy, Peer Leadership and Cultural Tolerance) the internal consistency is acceptable with values above .60. These results suggest that participants' scores are fairly consistent across the items that compose an individual scale. Previous Army research using the RBI (Knapp et al., 2005) also looked at the internal consistency of RBI scales when administering the full RBI assessment. The results found in Knapp et al. (2005) are consistent with the reliability results found in the current effort.

Along with examining the reliabilities for the RBI, correlations were calculated to determine the relationship between total score on the RBI and average supervisor ratings. Results failed to confirm a relationship between RBI total score and average supervisor ratings, $r = -.06$, *n.s.,* or overall effectiveness ratings, $r = -.15$, *n.s.* Additionally, there was no relationship between the RBI's Cultural Tolerance scale and supervisor ratings of "Exhibits Tolerance," $r = -.16$, *n.s.* A final correlation was examined based on similarities between the content of the RBI Diplomacy items and the definition of Supports Peers from the supervisor rating scales. Analysis shows that there is no relationship between the RBI measure of Diplomacy and supervisor ratings of "Supports Peers," $r = .02$, *n.s.*

The relationship between the RBI and other supervisor ratings dimensions are similarly low. Results show that most of the correlations between each pairing of the RBI scores and supervisor ratings are near zero or are negative and are not significant, implying that scores on the RBI are not effective predictors of supervisor ratings of performance. This finding is inconsistent with previous research (Knapp, McCloy, & Heffner, 2004; Knapp et al., 2005).

Previous Army selection and classification research examined convergent validity evidence for the RBI by examining the correlation between scores on the RBI and the International Personality Item Pool (IPIP), which is a measure of Big Five personality factors. A similar approach was taken in the current research. Participants were asked to complete the IPIP measure along with the other AISA tests. Knapp et al. (2005) hypothesized a relationship between the RBI Peer Leadership scale and Extraversion, the RBI Diplomacy scale and Extraversion and the RBI Cultural Tolerance scale and Agreeableness. They found statistically significant relationships in each of these hypothesized relationships. We evaluated the results of the current effort to examine these hypotheses. The results of these correlations are presented in Table 24. Results indicate that for Diplomacy and Peer Leadership the RBI has convergent validity as hypothesized in relation to the IPIP measure, however the hypothesized relationship between Cultural Tolerance and Agreeableness supported by Knapp et al. (2005) was not reflected in the current data.

Table 24. Correlations between RBI Scales and IPIP Constructs

RBI Constructs	IPIP Constructs	
	Extraversion	Agreeableness
Diplomacy	.26**	.12
Peer Leadership	.26**	.16
Cultural Tolerance	.23*	.14

[a]$n = 81$
* indicates correlation is significant at .05 level.
** indicates correlation is significant at .01 level.

Summary and Recommendations

The conflicting validation results between the RBI and the IPIP and between the RBI supervisor ratings indicate that additional research is required. The convergent validity evidence provided by the correlation between RBI scales and the IPIP measure is supported by the previous research programs in which the RBI was used as a predictor for selection and assignment (Knapp et al., 2004; Knapp et al., 2005). Given these results future investigations should be focused on exploring the criterion measures utilized in the current effort to ensure the ratings obtained are representative of the targeted dimensions.

Chapter 7: Semi-Structured Interview

Interviews are traditionally a good method for assessing soft skills such as those of interest in this effort (Pulakos et al., 1994). They generally show strong validity and low subgroup differences. These positive aspects outweighed the costs inherent in the personnel time required to administer and score interviews. The interview was "semi-structured" in that the item pool includes multiple questions that can be asked for each KSA. The interviewers could select the questions they want to ask in a given session. The semi-structured interview used a standard protocol for conducting the interview, selecting questions from a question bank, and evaluating interviewees in several target areas.

Instrument Development and Pilot

The first task in developing the interview was to determine which of the AISA dimensions lend themselves to an interview. One of our first decisions was to organize the questions into the higher-order KSAs shown in Figure 1(e.g., Peer Leadership) because we felt that scores at that level would be more stable than scores for the lower-level dimensions (e.g., acts as a role model, helping others, task leadership).

We drafted questions for most of the KSAs; however upon review, we found that in some instances, only one or two questions could be asked in a dimension (e.g., Dependability) without becoming redundant. This might ordinarily have been satisfactory, but, as we found when reviewing questions with Soldiers, there was essentially one answer to a question that asked how you demonstrate that you are dependable – "Do whatever it takes to get the job done." Since there was little variance in answers to Dependability questions, we dropped the dimension. We pilot tested some questions for Social Perceptiveness, but the pilot test demonstrated these questions were difficult to answer so we dropped that KSA.

The item pool originally contained items to assess Oral Communication (e.g., "What are some of the ways you have used to communicate technical or job information to people with differing levels of expertise?"). These questions did not seem to work well, so we dropped them because we wanted to ensure sufficient time to ask questions about the other KSA areas. Furthermore, we felt that it was not necessary to ask such questions when the interview itself offered a very good view of performance in that KSA. This approach is consistent with previous ARI research (Knapp et al., 2002).

The final set of interpersonal dimensions assessed in the interview was:

- Relating to and Supporting Others
- Conflict Management
- Cultural Tolerance
- Teamwork
- Adaptability /Flexibility
- Oral Communication
- Peer Leadership

After experimenting with several types of questions, we decided to use only experience-based questions for the AISA semi-structured interview. We considered including hypothetical questions, but found they elicited answers that were very similar in terms of the range of behavior. Everyone "knew" what a right answer would be and gave answers that indicated high performance. Experience-based questions ask the respondent to describe a) the situation in enough detail that the interviewer can get an idea of the challenges or difficulty in the situation, b) what action he or she specifically took (as opposed to what action a team took), and c) the outcome of the situation and action.

We originally developed four to ten items per dimension, with the expectation that many of them likely would be dropped during the SME reviews and pilot test. We pilot tested the semi-structured interview with volunteers from HumRRO and ARI. Four interviewer pairs met with each volunteer to ask a subset of the questions in the item pool. They took detailed notes for each answer, specifically noting the situation, action, and response. Interviewers took note of questions that respondents had difficulty answering. After the pilot test, project staff reviewed the interviewer notes and made the decision to drop questions that seemed to elicit the same kind of response or items that volunteers had difficulty in answering.

We presented this revised list of questions to NCOs (two females and four males) at Fort Jackson. They provided informal answers to the questions, and talked about the responses we were likely to receive from E3 and E4 Soldiers. These NCOs also helped us flesh out the rating scales, identifying behaviors we should look for at the three anchors – Exceeds Expectations, Meets Expectations, Needs Improvement. We incorporated these recommendations into the rating scales used in the field test and validity research.

The final rating scales, a sample of which is shown in Figure 12, had five components: (1) the KSA and its definition (see Appendix A for the definitions), (2) the seven-point rating scale, (3) the performance level names (i.e., Low, Moderate, High), (4) a brief general description for each performance level, and (5) more specific examples of behavior for each level. The interviewers marked their answers, with comments, on a separate form.

Instrument Validation

Procedure

Interviewer training. The interviewers were given the questions and rating scales, along with a written description of the process prior to the data collection. They also received a short briefing prior to the beginning of the data collection. The briefing consisted of discussing the item selection process and using the rating scales. Interviewers were told that they should take notes as a Soldier answered the interview questions, paying particular attention to the situation, action and result. At the end of the interview, they rated each Soldier on each KSA.

Due to operational restraints, the training time for the interviewers was too short to allow the interviewers to be completely prepared for their task at the onset. We did have the opportunity to answer questions and discuss problems during breaks, so after the first day, the interviewers felt much more comfortable in their roles.

Conducting interviews. We set a fifteen-minute time limit on the interview sessions so we could interview 24 Soldiers in a four-hour period. To meet what we considered to be severe time constraints, we limited the number of questions that a Soldier could be asked. Two interviewers each asked one question per KSA to each Soldier and then used the rating process described previously. If the Soldier failed to provide sufficient information on which to base a rating, the interviewers asked follow-up questions to probe the Soldier for additional information. If the probe failed to produce a response sufficient for rating, interviewers could opt to select another question within the KSA. In addition to asking questions and rating the Soldier responses, interviewers rated the Soldier's Oral Communication skills based on the Soldier's performance in the interview session. The ratings were averaged to create an interview score for each Soldier.

Cultural Tolerance		
The extent to which the Soldier demonstrated tolerance and understanding of individuals from other cultural and social backgrounds, both in the context of the diversity of US Army personnel and interactions with foreign nationals.		
1 **2**	**3** **4** **5**	**6** **7**
Low	**Moderate**	**High**
Soldiers low on this dimensions are not interested in learning about local cultures and do not worry about offending locals. They are impatient when working with people from varied cultures or backgrounds, refusing to take steps to overcome barriers. Soldier described situations in which s/he: • made or encouraged sexist, racist and culturally sensitive comments • made no effort to overcome language or cultural barriers	Soldiers moderate on this dimension are aware of the major aspects of local cultures, but may not always observe customs and so may sometimes offend locals. They are willing to work with people from varied cultures or backgrounds, but may accept barriers rather than trying to overcome them. Soldier described situations in which s/he: • did not make sexist, racist and culturally sensitive comments, but may have tolerated others doing so • made some attempt to overcome language or cultural barriers, but did not make strong efforts to do so	Soldiers high on this dimension find out about local cultures and learn customs so as not to offend locals. They enjoy working with people from varied cultures or backgrounds and are willing to work out solutions to overcome cultural barriers. Soldier described situations in which s/he: • took action to stop others from making sexist, racist or culturally sensitive comments • actively worked to overcome language or cultural barriers

Figure 12. Sample rating scale for semi-structured interview.

Results

Inter-rater reliability. The inter-rater reliability estimates for the interviewer ratings are presented in Table 25. The intraclass correlations (ICCs) were fairly low, which might be partially due to the fact that there were only two raters. The correlations between the raters are shown in Table 26; the numbers along the main diagonal are the interrater reliabilities for each dimension. There were some differences between the average ratings by the two raters. As shown in Table 27, with the exception of the Teamwork rating, Rater 2 provided higher average ratings than did Rater 1. Table 27 also shows the results of a paired-sample t test. The raters had

significantly different ratings for Relate to and Support Others, Conflict Management, Cultural Tolerance, and Peer Leadership.

Table 25 Inter-rater Reliability Estimates

Interview Dimension	ICC
Relating to and Supporting Others	.56
Conflict Management	.40
Cultural Tolerance	.63
Teamwork	.55
Adaptability/Flexibility	.41
Oral Communication	.67
Peer Leadership	.44

Descriptive statistics. The means for each scale were all close to 5.0, indicating moderately high average ratings (see Table 27). The standard deviations are relatively high for a 7-point scale, indicating that raters were seeing differences between people. The same two raters conducted interviews for all but one four-hour session, when another member of the data collection team substituted. The ratings for the interview scales were highly correlated with each other (as shown in Table 28) ($p < .01$). Table 29 shows the descriptive statistics for the interview scales. We could not test for subgroup differences gender or race/ethnicity due to the small sample size.

The correlations between performance ratings and interview ratings were close to zero (see Table 30) and frequently negative. The lone significant correlation can be attributed to chance. "Overall Effectiveness" is the final summary rating provided by supervisors; the scale asks for a global effectiveness rating rather than one related to performance in a specific area. "Average Rating" is the average of the 12 rating dimensions, without the overall rating.

Table 26. Intercorrelations of Interviewer Ratings

	Rater 2						
Rater 1	Relate to Others	Conflict Management	Cultural Tolerance	Teamwork	Adaptation to Change	Oral Comm	Peer Leadership
Relate to Others	.56**						
Conflict Management	.42**	.40**					
Cultural Tolerance	.41**	.32**	.63**				
Teamwork	.52**	.37**	.48**	.55**			
Adaptation to Change	.41**	.29*	.42**	.42**	.41**		
Oral Communication	.52**	.41**	.41**	.60**	.51**	.67**	.
Peer Leadership	.46 **	.36**	.30**	.38**	.40**	.59**	.44**

** Correlation is significant at the 0.01 level (2-tailed).
 * Correlation is significant at the 0.05 level (2-tailed)

Table 27. Average Interviewer Rating, Mean Difference, and T-test Results

	Rater 1		Rater 2		Mean Difference	Std. Dev.	t	df	p
	Mean	Std. Dev.	Mean	Std. Dev.					
Relate to and Support Others	4.63	1.55	4.96	1.51	-.29	1.43	-1.97	91	.05*
Conflict Management	4.23	1.56	4.81	1.64	-.54	1.76	-2.96	91	.00**
Cultural Tolerance	4.54	1.25	5.01	1.50	-.43	1.21	-3.34	88	.00**
Teamwork	5.09	1.21	4.81	1.17	.20	1.14	1.61	84	.11
Adaptation to Change	4.72	1.21	5.00	1.29	-.25	1.36	-1.76	92	.08
Oral Communication	4.66	1.58	4.78	1.55	-.09	1.27	-0.67	89	.51
Peer Leadership	4.51	1.56	5.19	1.55	-.63	1.63	-3.74	93	.00**

** Correlation is significant at the 0.01 level (2-tailed).
 * Correlation is significant at the 0.05 level (2-tailed)

Table 28. Correlations Between Interview Dimension Scores

	Relate to Others	Conflict Management	Cultural Tolerance	Teamwork	Adaptation to Change	Oral Communication
1 Relate to Others						
2 Conflict Management	.56**					
3 Cultural Tolerance	.53**	.55**				
4 Teamwork	.63**	.59**	.60**			
5 Adaptation to Change	.69**	.65**	.57**	.72**		
6 Oral Communication	.70**	.66**	.61**	.73**	.73**	
7 Peer Leadership	.63**	.60**	.47**	.56**	.63**	.76**

** Correlation is significant at the 0.01 level (2-tailed).

Table 29. Average Ratings and Standard Deviations across Raters for Interview Scales

Rating Dimension	N	Min	Max	Mean	SD
Relate to Others	96	2	7	4.77	1.373
Conflict Management	96	1	7	4.52	1.347
Cultural Tolerance	95	1	7	4.74	1.263
Teamwork	95	1	7	4.99	1.053
Adaptation to Change	96	2	7	4.84	1.072
Oral Communication	96	1	7	4.72	1.434
Peer Leadership	96	2	7	4.82	1.359

Note: One Soldier was not rated because he did not speak fluent English; two other Soldiers were not able to answer the interview questions.

Table 30. Correlation between Interview Ratings and Performance Ratings

	Supervisor Ratings			
Interview Rating	Overall Effectiveness	Average Rating	Interpersonal Scales	Teamwork
Relate to and Support Others	-.01	.02	.12	.17
Conflict Management	-.07	-.01	.06	.04
Cultural Tolerance	-.01	-.01	.03	.01
Teamwork	-.02	.00	.07	.13
Adaptability	-.10	-.02	.06	.12
Oral Communication	-.08	-.01	.05	.11
Peer Leadership	-.07	-.13	.01	.04

Note. * *Significant at the 0.05 level*

Summary and Recommendations

These results are very disappointing. However, given the past research that has demonstrated the validity of structured interviews, we must examine the circumstances in the current effort that might have caused these unexpected results. In discussing the preliminary results with ARI staff, we speculated about several possible explanations for these results.

First, respondents typically take several minutes to answer each question in a structured interview. These respondents gave responses that were generally short and to the point, with little elaboration. This might be an indication that they were reluctant to talk in detail about situations that might have been uncomfortable. In addition, these Soldiers had returned from Iraq only a few months before the data collection. Compared to their deployment experiences, which still might be affecting these Soldiers, this exercise might have seemed pointless to some and threatening to others.

The U.S. is currently at war on several fronts, resulting in large scale troop deployments. This has meant that requests for troop support were either trimmed severely or denied altogether. If not for this large-scale deployment, we would likely have been able to collect data in multiple locations and obtain a larger sample size. The sample size alone limits the strength of the conclusions at which we can arrive based on these data.

However, we cannot put the blame on our sample for all the shortcomings with this instrument. Among the main points we should consider are these:

- We had several decision points along the development process, most of which resulted in reducing the number of questions. We should review dropped items to see if they can be salvaged. We dropped many questions from the bank because they were redundant or because they failed to elicit responses in the field test. We likely put too much weight on the field test. We have used these questions or some that

were very similar in other projects and had good success. We should review those dropped items, being more liberal in our decisions to drop or keep them. In some cases, the redundant questions could be used when an interviewee has difficulty answering a particular question. Also, retaining some of those dropped items will give interviewers more options to use if an interviewee is unable to answer a specific question.

- Operational constraints prohibited interviewer training. Typically, rater training includes practice in asking the questions, making ratings, and in the use of probes to help an individual answer a question. Improving rater training and their understanding of their roles would likely improve the outcomes.

- Had this project gone into Phase III, we would have undertaken a comprehensive training program for military raters. These raters would be more familiar with Soldier jobs and situations and might be more likely to elicit the type of information we wanted. Military raters would also be more likely than our staff raters to understand the implications of what the Soldiers said.

- Typically, we asked raters to reach consensus (i.e., to be within one point of each other) to determine a final rating. However, we did not follow that procedure in this research because we were concerned that it would be too time consuming and would prevent us from collecting interview data from all participating Soldiers. Future administrations should include time for raters to reach consensus on their ratings. The consensus discussion helps them identify common themes and calibrate themselves to each other. This is a technique commonly used in assessment centers, and has worked well with interviews in previous research.

- This concern about time might also have led the interviewers to be more willing to move to a new question rather than waiting for a Soldier to think for a few minutes to formulate an answer. If we were to collect additional data or actually administer the AISA battery, we would allow additional time for the interview (15 to 20 minutes for questions, 5 minutes for ratings). We anticipated difficulty with completing the interviews in the allotted time; however, this did not turn out to be the case. It might have been better to collect quality data from fewer Soldiers than to be concerned about including all possible Soldiers. This tradeoff between quality and quantity is a constant struggle in research efforts such as this.

Chapter 8: Leaderless Group Discussion (LGD) Exercises

Like the interview, the two separate LGD exercises were developed to assess Soldier skills. In the first LGD, a four-person group was tasked to help a family plan a day-long tour of Washington, D.C. In the second exercise, the group was tasked to help a town determine several factors related to building a community center. Both exercises were adapted from a master's thesis (Brockson, 1999). While planning a day tour and determining a community center location are not typical Soldier tasks, they could still be used to assess Soldier interpersonal skills. Recall that the purpose of the exercises was to provide a stimulus to which participants would react and demonstrate their interpersonal skills. It was not critical to develop Army specific exercises as long as Soldiers could relate to them, whether as part of their professional or personal lives. This chapter presents the development activities, field test, and validation results for both exercises.

Instrument Development and Pilot Test

Development of the LGD instruments was an iterative process that required several SME reviews. First, we created a draft of the two exercises designed to create situations that would provide participants with the opportunity to exhibit interpersonal behaviors. Cultural Tolerance was dropped from this list because we found it could not be assessed easily in the exercises. The LGD tasks (i.e., planning a day tour, building a community center) were neutral in nature and did not elicit this behavior from participants. In the exercises, participants could potentially find themselves in a culturally diverse group and might need to display cultural tolerance. However, this would be dependent on group composition and would vary by group. After eliminating Cultural Tolerance, we expected to be able to evaluate participants' performance on these six interpersonal dimensions:

- Relating to and Supporting Others,
- Conflict Management,
- Teamwork,
- Adaptability/Flexibility,
- Communication, and
- Peer Leadership.

We held two instrument development workshops where senior NCOs reviewed the LGDs. Then we pilot tested and field tested the instruments. At the first workshop, NCOs reviewed a set of participant materials (i.e., the instructions, scenario description, maps) for clarity and appropriateness for E3-E4 Soldiers. They were asked to comment on the reading level, complexity, and length of each packet, and to provide feedback about whether there was sufficient information about each aspect of the scenario. We were also interested in their general reactions to the exercises and the estimate of time it would take to review the materials provided for the exercises.

We next administered the LGD exercises to four-person NCO groups who went through the exercises, read the materials and conducted a discussion. NCOs were given 25-30 minutes to review the LGD materials, 30 minutes to discuss the problem as a group, and a few minutes to summarize their recommendations. After completing the LGDs, Soldiers commented on their

47

general reactions to the exercise, provided feedback about how well they thought the exercises tapped into the targeted KSAs, and how they thought the target population, E3-E4s, would respond to the exercise (e.g., whether they would use the same problem-solving strategies). The administrator took notes and audio taped the LGD discussions. The exercises generally received positive reactions. The NCOs indicated that the DC Tour would probably be easier for Soldiers because the task was similar to something that might be done in the Army. They had some concerns that Soldiers would not relate to the problem posed in the Community Center, although they said Soldiers would be able to complete the exercise.

We pilot tested the LGDs with eight E3 and E4 Soldiers – two groups of four participated in both exercises. Soldiers read the instructions, reviewed their participant materials, and then took part in the group discussion. After the discussion, the pilot test administrators led a discussion of the exercise with the participants, focusing on changes that might improve the exercise. Both the group discussion and administrator-led discussion were videotaped. As a result of the feedback received in the pilot test, we made changes to clarify the initial instructions and to provide more structure (i.e., reduce the number of solution paths) to the Community Center exercise.

The LGD instruments were field tested at Fort Riley. As a result, we significantly revised the rating procedure from a performance rating scale and checklist hybrid to a simple checklist. We also asked participants to rate their own and each other's performance during the discussion. It was easy for one person to dominate the discussion in an LGD, which was not desirable from our perspective. To encourage everyone to participate, we informed them at the beginning of the exercise that they would be rating themselves and each other on how well they exhibited the six interpersonal dimensions during the discussion. Only minor revisions were made to the exercises after field testing. We made editorial changes to the materials and reorganized the participant packet materials so that they would be more convenient to use.

Specific development activities for each exercise and changes made as a result of feedback received from the site visits are discussed in more detail in the following sections. The general procedure for both exercises was the same. Each participant had a packet of materials that included background information about the problem and instructions for completing the task. The test administrator read the instructions to the Soldiers, who read along from their own packets. Participants had ten minutes to read the information relevant to their task, and then had thirty minutes to discuss the situation and come to agreement on an answer. One member of the group volunteered or was designated by the group to report the results to the observers.

DC Tour Exercise

In the DC Tour exercise, participants were tasked to plan a tour of Washington, DC for the Jones family. The Jones family was in town for only one day and wanted to visit all the attractions on their list (see Figure 13). They obtained information about each location, but still needed help creating an itinerary for the day. Participants worked as a team to plan a schedule that would allow the family to visit all locations and return to the point of origin by a given time. The DC Tour participant materials contained (a) an introduction and instructions for completing the task, (b) additional information about each tour location, (c) a calendar for creating the day's schedule, and (d) a map of the National Mall with important features labeled.

48

Mr. Jones: Vietnam Veterans Memorial, Museum of American History, WWII Memorial

Mrs. Jones: Museum of American History, Washington Monument, Lincoln Memorial

Junior: Zoo, Natural History Museum, Air and Space Museum, White House

Figure 13. Sites the Jones family wanted to see on their tour.

Participants received one of four participant packets. The packets contained the same basic information about all nine locations; however, each packet contained slightly different information from the rest. Moreover, no one person had *all* the information needed to create the best itinerary. The information varied in specific information about (a) the time the family would spend at each location, (b) scheduling restrictions (e.g., museum closes at 2:00 pm), (c) hours of operation, and (d) travel time between locations. The materials also contained some irrelevant information. The intent was for participants to discover they had different pieces of information and to combine their knowledge to arrive at a solution. These pieces of information were distributed evenly across participants.

Participants were assigned three types of information. Common information, such as knowing that it took 15 minutes to visit the World War II Memorial, was accessible to all participants. Partially shared information was known to two individuals and was important for planning, but would not have a large impact on the outcome if undiscovered. Unique information was given to only one participant, and without it the team would not be able to find the optimal solution. The developers created a table to help allocate task information equally between participants.

Twenty minutes into the group discussion, the Test Administrator distributed a critical piece of information to each participant and told them they must incorporate it into their schedule. This information indicated the family needed to obtain tickets to tour the Washington monument and tickets were only sold once in the morning and once in the afternoon. Additionally, the line in the afternoon was longer and thus would take more time to maneuver. Each participant received the same critical information and he or she was to incorporate it in the group planning, revising the group's tour schedule as necessary. The introduction of this new material was intended to capture participants' ability to adapt to changing situations, should they decide to do so. The new information was critical because it required participants to rework their tour schedule to allow time for obtaining tickets to the Washington monument. This must be done to reach the optimal solution. However, participants ignored this information occasionally and did not incorporate it in their planning.

Community Center

In the Community Center exercise, four participants were tasked with helping a fictional town, Lampsburg, evaluate proposals for building a community recreation center. Participants played the role of members of an independent commission appointed by the town council to review summaries of three proposals for the center and to make recommendations as to how the town should proceed. Participants were informed that members of the town council disagreed on three areas of the proposals: (a) the features or facilities the center should have, (b) where the

center should be located, and (c) where to obtain funding for a down payment for the center. Additionally, participants were told that the town council offered them specific guidance (e.g., sources of funding, total cost of the center) to which they must adhere. The funding sources provided different amounts of capital and usually had restrictions attached (e.g., must build in a specific location).

The participant materials for the Community Center exercise consisted of (a) a description of background information about Lampsburg (e.g., history, size, community values), (b) a map of the town that showed each of the proposed sites, and (c) summaries of each of the three proposals. Each proposal's summary was broken down into sections that provided details about features of the proposed center (e.g., weight rooms, parking), the proposed location of the center, sources of funding for down payment, and public comments on the proposed center.

The administrator read through the introduction with the participants, then gave them 15 minutes to review the packets of materials. After the review, participants took part in a 30-minute group discussion. The administrator informed them that by the end of their discussion they would have to agree on a set of recommendations to make to the council and be prepared to describe the rationale behind those recommendations.

Like the DC Tour exercise, the Community Center exercise also contained varying amounts of unique, partial, and shared information regarding the proposals that each participant received. In this case, all the different information was included in comments from individuals and citizen groups who provided their opinions on key elements of the proposals (e.g., facilities, location, sources of funding). The comments were balanced so there were equal amounts of information that made one proposal look more attractive than another.

LGD Evaluation Instruments

We evaluated participants' performance on six interpersonal dimensions—Relating to and Supporting Others, Conflict Management, Teamwork, Adaptability/Flexibility, Oral Communication, and Peer Leadership. Our expectation was that the AISA battery would be used to help Soldiers identify developmental needs. So, we decided to take advantage of the innovative aspects of the project to try out a new rating system. We developed a scoring instrument that combined behaviorally anchored rating scales with an embedded checklist, with the idea that the rating scale would allow us to provide a numeric rating and the checklist would allow us to provide specific, qualitative feedback to participants. However, the instrument proved to be too cumbersome to use.

As a result, after the field test, we developed a checklist that listed specific positive and negative behaviors for each dimension (see Figure 14 for an example) and that could be easily used to capture participant behavior and could provide both a numeric rating and qualitative feedback. Negative behaviors were indicated by italics. We wanted to be able to distinguish between a one-time behavior and those that occurred repeatedly. Two researchers who served as raters checked one box next to the behavior the first time it was observed. If a behavior was demonstrated more than once or intensely, the raters indicated so by checking both boxes. The raters completed a different checklist for each exercise. The simplified checklist was much easier to use.

We also created and administered a peer feedback instrument to participants as an incentive to participate. Although this was also an opportunity to gain additional data, the primary reason for including it was to apply some peer pressure that might motivate Soldiers to take part in the discussion. Though it was not a big incentive, we believed if we told Soldiers there would be some consequence to their actions, (i.e., performance would be evaluated by their peers), they would put forth more effort in the exercise than if there was no consequence at all. Participants were instructed at the beginning of the exercise that after the discussion period they would rate themselves and their teammates on their performance during the exercise. The peer feedback instrument listed the same six dimensions that are on the behavioral checklist and employed a 5-point rating scale. Space was provided for Soldiers to rate up to four participants. Figure 15 shows the instructions and two of the rating scales used in the instrument. This instrument also provided an extra set of data to use in the analysis.

Instrument Validation

Procedure

As described earlier, the validation testing occurred in four classrooms. One room each was used to administer the computer-based tests, DC Tour exercise, Community Center exercise, and Semi-Structured Interview. Participants represented various MOS, with the majority of them from the 11B or 88M MOS (refer to Table 4). All Soldiers began testing in the computer-based room, and the computer room administrator randomly assigned them to their next exercise. To control for group familiarity in the LGD exercises, the administrator randomly assigned four Soldiers to take part in those exercises. As Soldiers reported to the LGD rooms, they took a seat at the table and received one of the four participant packets. To facilitate discussion, we arranged the tables so that two participants sat on each side. The administrator read the exercise introduction and instructions and gave them 10 minutes (15 minutes for the Community Center) to review the materials. As Soldiers began the group discussion, two raters (i.e., the researchers) used the behavioral checklist to evaluate the live performance. To make the rating process manageable, the plan was for each rater to focus on two participants. However, the checklist instrument was very easy to use and for the most part, observing and evaluating four participants at a time was easily accomplished. After 30 minutes, the group was given a few minutes to summarize its recommendation to the administrators. Finally, participants completed the peer feedback instrument. All exercises were videotaped to allow the developers to see how the process worked.

Results

LGD behavioral checklist. We ran inter-rater and intra-rater correlations on the checklist data to assess the level of agreement between raters (interrater reliability) and the extent to which each rater was consistent in his or her ratings on each scale (intra-rater reliability). We then conducted an analysis that took both types of reliability, interrater and intra-rater, into account. The results, shown in Table 31, indicated that most of the reliabilities were not acceptable. The reliability indices of all Community Center scales were greater than .30, with most of them being .45 or greater. The reliabilities of the DC Tour scales were generally higher than those of the Community Center, although none reached generally accepted levels of reliability.

```
┌─────────────────────────────────────────────────────────────────────┐
│                             Instructions                            │
│                                                                     │
│ Using the checklist on the next page, mark all the behaviors that you observed from the │
│ participant during his/her interactions with other participants on the exercise. │
│                                                                     │
│  Check one box for each behavior that occurred once, or was *not* demonstrated repeatedly or │
│  intensely.                                                         │
│                                                                     │
│  ❑    Use two boxes when the behavior in question was observed repeatedly or was used │
│       intensely.                                                    │
│                                                                     │
│  *Note that negative behaviors are shown in italics at the end of each KSA.* │
│                                                                     │
│                                                                     │
│                    Relating to and Supporting Others                │
│                                                                     │
│ ❑    ❑    Was courteous and respectful to others                    │
│ ❑    ❑    When disagreed with someone, did so politely              │
│ ❑    ❑    Helped others understand specific points or rationales    │
│ ❑    ❑    *Was rude or disrespectful to others*                     │
│                         Conflict Management                         │
│ ❑    ❑    Clarified when others misunderstood each other's points   │
│                                                                     │
│ ❑    ❑    Proposed tradeoffs or compromises                         │
│                                                                     │
│ ❑    ❑    Responded freely and politely when others questioned or disagreed with Soldier's ideas or │
│           suggestions                                               │
│                                                                     │
│ ❑    ❑    *Refused to listen to others' viewpoints; insisted on own solution* │
│                                                                     │
│ ❑    ❑    *Became defensive or angry when others questioned or disagreed with Soldiers' views or* │
│           *suggestions*                                             │
└─────────────────────────────────────────────────────────────────────┘
```

Figure 14. Instructions and example from LGD checklist.

Peer ratings. The LGD checklist dimensions correlated highly with the peer rating dimensions, particularly for the DC Tour exercise (see Table 32). Many of the correlations were significant, $p < .01$. These findings indicated that observers evaluated participants consistent with the way participants rated themselves and each other. These results suggested that the checklist was a viable instrument for measuring a particular construct.

Group Exercise Rating Scales

Date: _____

We want to know how you think **_you_** **_and_** **_the_** **_other members_** of your team performed on this exercise. Please carefully follow the three steps described below:

1. Write the ID # of all Soldiers in your group in the order you will rate them, starting on the far left side of the group. Put an **X** next to your number.

> Soldier 1 _____
>
> Soldier 2 _____
>
> Soldier 3 _____
>
> Soldier 4 _____

2. Check the exercise you are doing _____ DC Tour _____Community Center

3. Rate yourself and the other members of your group on each performance area.
   a. Read the definition of each area.
   b. For each area, rate each group member – including yourself – from 1 (poor) to 5 (excellent) on how well you think they performed in the exercise.

Relating to and Supporting Others

The degree to which an individual treats others in a courteous, respectful and tactful manner; provides help and assistance to others; is sensitive to others' priorities, interests, and values; and exhibits good will towards others and is tactful and helpful.

	Needs Improvement		Adequate		Strength
Participant 1	①	②	③	④	⑤
Participant 2	①	②	③	④	⑤
Participant 3	①	②	③	④	⑤
Participant 4	①	②	③	④	⑤

Conflict Management

The degree to which an individual encourages and supports different perspectives; avoids harmful conflict; constructively addresses disagreements that undermine group performance; and deals with conflicts in ways that preserve good relations and enhance trust.

	Needs Improvement		Adequate		Strength
Participant 1	①	②	③	④	⑤
Participant 2	①	②	③	④	⑤
Participant 3	①	②	③	④	⑤
Participant 4	①	②	③	④	⑤

Figure 15. Example peer feedback instrument.

Table 31. Scale Psychometric Properties for Community Center and DC Tour

	Combined Score					
	Total # of Items in Sub-scale	# of Reverse scored items	Inter-Rater Rel.	Rater 1 Intra-Rater Rel.[1]	Rater 2 Intra-Rater Rel.[2]	Reliability[3]
Community Center[4]						
Relating to Others	4	1	.35	.39	.67	.45
Conflict Management	5	2	.21	.25	.50	.34
Teamwork	5	1	.46	.49	.55	.53
Adaptability	3	0	.21	.66	.55	.31
Communication	5	0	.37	.52	.69	.54
Peer-Leadership	8	2	.75	.70	.76	.63
DC Tour[5]						
Relating to Others	4	1	.37	.57	.44	.26
Conflict Management	5	2	.29	.53	.61	.22
Teamwork	5	1	.59	.76	.60	.64
Adaptability	5	2	.51	.42	.62	.54
Communication	5	0	.58	.86	.72	.71
Peer Leadership	8	2	.81	.80	.81	.81

[1] These are the internal consistency of the scales. [2] These are the internal consistency of the scales. [3] These reliabilities are coefficients that account for both item and rater specific factors as sources of measurement error. [4] $n = 70$ [5] $n = 68$.

Supervisor ratings. Analyses showed that most of the correlations between the LGD checklist (observer ratings) and supervisor ratings were near zero or negative. Correlations near zero implied that scores on the LGD checklist did not predict supervisor ratings of performance, and significant negative correlations suggested that there was an inverse relationship between the scores on the LGD checklist and supervisor ratings. That is, high performance on the LGD predicted low supervisor ratings. These findings are presented in ***Table 33***. These findings are contradictory to the acceptable relationships between the peer and observer ratings. Two possible reasons exist for these differences. First, the supervisors were rating overall behaviors whereas the peers and observers were rating behaviors in a very specific situation. This explanation is reinforced by the differences found between the two LGD tasks, suggesting that the context does play a role in rating agreement. Second, a Soldier's peers likely interact with their peers in a broader range of situations, e.g., off duty environments, than a supervisor. This broader knowledge of their peer interaction behaviors may have impacted the peer ratings.

Subgroup differences. Because of the small sample size, we could not calculate subgroup differences (e.g., by gender, race, MOS). Although the results may not be stable, it may be noteworthy that scores for E4 Soldiers were consistently higher than scores for Soldiers in lower pay grades. Perhaps this implies that Soldiers become more interpersonally skilled with experience and age.

Table 32. Correlations between Checklist Scale and Composite Scores and Peer Ratings

Checklist Sales and Composites	Community Center Peer Rating						DC Tour Peer Rating					
	1	2	3	4	5	6	7	8	9	10	11	12
Community Center												
1. Relate to Others	.49***	.27*	.44***	.15	.56***	.45***	.33*	.17	.12	.09	.23	.20
2. Conflict Management	.60***	.48***	.46***	.56***	.44***	.28	.32*	.24	.27	.23	.21	.26
3. Teamwork	.29***	.15	.36***	.07	.49***	.42***	.05	.03	.06	-.08	.15	.11
4. Adaptability	.14	.03	.18	.04	.27*	.21	.12	.08	-.02	-.07	.24	.08
5. Peer Leadership	.37**	.23	.37***	.12	.45***	.39***	.19	.14	.20	.15	.32*	.21
6. Communication	.32***	.00	.22	.04	.38***	.37***	.16	.01	-.02	-.02	.16	.03
DC Tour												
7. Relate to Others	.28*	.26	.23	.15	.13	-.10	.45***	.38***	.43***	.40***	.39***	.46***
8. Conflict Management	-.06	-.01	.06	-.03	-.10	-.16	.25*	.21	.20	.19	.25*	.24*
9. Teamwork	.26	.26	.15	.06	.25	.11	.59***	.55***	.55***	.52***	.57***	.58***
10. Adaptability	.21	.19	.20	.16	.15	.05	.22	.31***	.32***	.19	.27*	.37***
11. Peer Leadership	.27*	.23	.26	.09	.33*	.31*	.62***	.56***	.58***	.48***	.66***	.60***
12. Communication	.32*	.33*	.26	.06	.35*	.21	.56***	.51***	.50***	.44***	.59***	.58***

Note. n $_{CC-CC}$ = 70; n $_{DC-DC}$=68; n $_{CC-DC}$=52. * = p<.10; ** = p<.05; ***= p<.01. Peer ratings were averaged across 3 raters.

Table 33. Correlations between Checklist Scores and Supervisor Ratings

Check-List Subscales and Composites	Performance Rating Dimensions				
	Communication	Adaptation	Support Peers	Tolerance	Overall
Community Center					
Relate to Others	.03	-.27**	-.12	-.07	-.24
Conflict Management	-.07	.02	-.11	.10	-.06
Teamwork	-.17	-.14	-.03	-.16	-.09
Adaptability	-.10	-.12	-.06	-.17	-.07
Peer Leadership	-.08	-.22*	-.10	.00	-.15
Communication	-.04	-.40***	-.12	.05	-.40***
DC Tour					
Relate to Others	-.02	.07	.07	.12	.04
Conflict Management	-.01	.09	.04	.05	.07
Teamwork	.00	-.10	-.17	.17	-.22*
Adaptability	-.17	-.11	-.24*	-.14	-.21
Peer Leadership	.05	-.13	-.27**	-.04	-.19
Communication	-.08	-.15	-.19	.01	-.22

Note. N = 55 - 57 * = *p*<.10; **= *p*<.05 ; ***= *p*<.01

Summary and Recommendations

Taken as a whole, we believe the LGD instruments have potential to be a viable tool for measuring interpersonal skills and would add value to an organization's assessment system. The technique engaged the Soldiers and elicited the types of behavior that represent interpersonal skills. Further, despite the lack of interrater consistency, many of the relationships between the observer and peer ratings were acceptable. However, before the LGD is implemented operationally, we would recommend several changes to the instrument (i.e., the LGD exercise and checklist) and potential application for each exercise.

Revisions in the LGD Exercise

After multiple administrations, it became apparent that the Community Center exercise may have imposed a higher cognitive load on participants than the DC Tour exercise. The Community Center presents a large amount of information that must be read and absorbed in a short amount of time. We offer two recommendations to make the exercise less difficult. The first is to simplify the sentence structure and eliminate extraneous information in the participant packets. While having some extraneous information in the exercise is desirable, we believe the Community Center may have contained too much information (more so than the DC Tour exercise). Second, we suggest reducing the number of features that could be added or the sources of funding, so that there are fewer factors to consider. Reducing the amount of reading and information to process may lead participants to be more successful in this exercise.

Revisions in Rating the Exercise

Results from the validation effort provided mixed support for the LGD checklist as a good instrument for predicting interpersonal skills. The weak relationship between the LGD checklist and supervisor performance ratings could be attributed to several factors, one of which was a small sample size ($n < 70$). Another explanation could be that the checklist instrument, while easy to use, did not adequately capture the range of behaviors that participants demonstrated. Alternatively, the checklist may be a good measurement method, but the observer training may have been insufficient.

The checklist allowed the observer to capture specific behavior to use for feedback, which was the intention for the original rating scales. We recommend continuing use of the checklist, with modifications. To make the checklist more accurate, we recommend conducting SME workshops with the anticipated users (e.g., military, civilians) to help identify the most important behaviors to include in the checklist. Some of the behaviors (e.g., use of profanity) which we consider negative, might be completely acceptable to the users. This is especially true in the military domain.

We also recommend supplementing the checklist with rating scales that would allow a useful summary of performance and providing feedback to participants. The scales used for the Semi-Structured Interview could be adapted for this exercise. Observers would use the checklist during the discussion period, make individual ratings while participants are completing their peer ratings, and then reach a consensus rating after the participants leave the room. Participants would receive a score based on the rating scale and could also receive either a copy of their checklist or a more formal summary based on the checklist. If our goal is to help Soldiers improve their interpersonal performance, we should give them the most specific feedback we can.

Application of the LGD Exercises

During the concurrent validation, we found that the DC Tour and Community Center exercises elicited different behaviors from participants. The DC Tour is a problem-solving task that focuses on facts and a strict timeline (i.e., scheduling activities), activities with which the participants were very familiar. Because participants were more familiar with the task presented in the DC Tour, they spent little time figuring out what to do. Also, because the exercise was very concrete there was little reason to debate with one another and they could focus on working as a team.

The Community Center exercise is more abstract and opinion based (i.e., advocate one's position for the recreation center), and presents participants with a situation in which they may not be used to working. It requires at least one participant to step up and guide the discussion. There are also more opportunities to disagree with one another.

Because the two exercises seem to measure different interpersonal competencies, they should be used for different purposes. For example, if the purpose of employing the LGD is to measure one's ability to work in a team setting and to adapt to changing situations, the DC Tour would be appropriate. Alternatively, the Community Center exercise would be more suitable in determining Conflict Management and Peer Leadership skills. We view the two exercises as different, but complementary to one another. Lastly, if these exercises are to be used in the military setting, the scenarios should be adapted further to reflect more Soldier-relevant tasks. The key is to choose a scenario that fits the target organization and the purpose of the assessment.

Chapter 9: Cross-Instrument Analyses

Along with exploring the properties of the individual assessments that comprise the AISA, we also analyzed the properties of the assessment battery as a whole. This chapter discusses the relationship among overall assessment scores on each test as well as the relationships between multiple measures of individual KSAs as measured by the various assessments. Additionally, in this section of the paper we discuss the hypothesized relationships between the individual KSA scores and the IPIP dimensions.

Individual KSA Scores

Adaptability/Flexibility

The AISA battery included three assessments that measured the Adaptability/Flexibility construct: the interview, the Community Center LGD, and the DC Tour LGD. The analysis found a significant positive relationship between Adaptability as measured by the interview Adaptability and the DC Tour exercise, $r = .24$, $p < .05$. No significant relationship was found between the interview and Community Center ($r = .09$, *n.s.*) or between the Community Center and DC Tour ($r = .14$, *n.s.*) scores for Adaptability.

Conflict Management

The Conflict Management KSA was measured by the interview, the SBISE and both LGD exercises. Table 34 shows the results of the analysis of the relationships between the multiple measures of conflict management in the AISA. The only significant relationship found was the relationship between the DC Tour measure of Conflict Management and the SBISE Conflict Management subscale ($r = .26$, $p < .05$).

Table 34. Correlation between AISA Measures of Conflict Management

	1	2	3
1. Interview: Conflict Management			
2. SBISE: Conflict Management	.14		
3. Community Center: Conflict Management	.00	-.05	
4. DC Tour: Conflict Management	-.04	.26*	.14

Note. * *Significant at the 0.05 level*

Cultural Tolerance

Cultural tolerance was measured by the interview, the SBISE and the RBI. A significant positive relationship was found between interview scores on cultural tolerance and on the cultural tolerance RBI subscale ($r = .22$, $p < .05$). There was not a significant relationship between the SBISE cultural tolerance scale and either the RBI ($r = -.17$, *n.s.*) or the interview measures ($r = .05$, *n.s.*) of cultural tolerance.

Peer Leadership

Peer Leadership is composed of three individual KSAs: Task Leadership, Acts as a Role Model, and Helping Others. Any assessment items aimed at one of these competencies was included in the Peer Leadership construct which is measured in the interview, the SBISE, the RBI, and both LGD exercises. Table 35 contains the results of the analysis of the relationship of the multiple measures of Peer Leadership. As seen below, several significant relationships exist between the measures of Peer Leadership, with only the SBISE measure not significantly related to the other measures.

Table 35. Correlation Between Peer Leadership Measures in the AISA Battery

	Interview	SBISE	RBI	CC
Interview: Peer Leadership				
SBISE: Task Leadership	-.02			
RBI: Peer Leadership	.28**	-.08		
Community Center: Peer Leadership	.25*	.15	.65**	
DC Tour: Peer Leadership	.25*	.20	.45**	.51**

Note. ** indicates significance at the .01 level; * *S*ignificant at the 0.05 level

Relating to and Supporting Others

Relating to and Supporting Others is comprised of three individual KSAs: Ability to Relate to and Support Peers, Amicability, and Concern for Soldier Quality of Life. The construct was measured by the interview, SBISE, and both LGD exercises. Table 36 contains the results of the analysis of the relationships between multiple measures of Relating to and Supporting Others. As seen below, significant relationships exist between the interview measures of Relating to and Supporting Others, but the SBISE measure was not significantly related to either of the others.

Table 36. Correlation between Relating to and Supporting Others Scales in the AISA Battery

	Interview	CC	DC
Interview: Relate to and Support Peers			
Community Center: Relate to and Support Peers	.26*		
DC Tour: Relate to and Support Peers	.25*	.38**	
SBISE: Relate to and Support Peers	-.10	.12	.06

Note. ** indicates significance at the .01 level; * *S*ignificant at the 0.05 level

Communication Ability

The communication ability construct is composed of both Oral and Written communication components. The written communications element was measured in the WCA while the oral communications element was gauged in both the interview and the LGD exercises. Because these components measure distinct abilities that can be combined to yield an overall picture of an individual's communication ability, it is reasonable to expect that WCA scores for written communications may not be related to the measures of oral communication ability as

measured in the LGD exercises and the interview. Analyses support the notion that the written and oral measures represent different constructs and therefore, the only relationships explored here are those between the measures of Oral Communication. The analysis indicates no significant relationships exist between the Community Center measure of Oral Communication and the interview, $r = .18$, *n.s.*, the DC Tour and interview, $r = .23$, *n.s.,* or the two LGD measures of Oral Communications ability, $r = .26$, *n.s.*

KSA Scores and IPIP Traits

In Chapter 1 of this report we outlined our belief that both general mental ability (GMA) and trait dispositions have effects on knowledge and demonstration of interpersonal skills. To test the relationship of trait dispositions with our measures of interpersonal skills we included the IPIP measure of the Big Five personality traits as a marker test, to show whether the AISA instruments assess the relevant personality measures. Table 37 shows the correlation of AISA measured KSAs and the IPIP scales. Underlined cells denote specific hypothesized relationships between the AISA KSA and the IPIP trait measure. Only the SBISE measure for Concern for Quality of Life and the IPIP measure of Agreeableness show a significant relationship in the hypothesized direction. Overall, there appears to be no relationship to performance on the AISA trait measures and interpersonal skills measures found in the IPIP.

Overall Test Scores

As discussed previously in this report, the supervisor ratings used as the criterion measure in the validation effort focused on overall performance rather than on interpersonal KSAs. The ratings appear to have been affected by halo error, which limits the extent to which they could correlate with the AISA interpersonal measures. Because of measurement error it would be premature to assert the criterion related validity of the individual construct measures for use in selection and assignment decisions. This led us to report an overall score for each assessment to be used in a selection and assignment context rather than by KSA. The statistical properties of those overall scores and their relationships are described in this section of the report. An overall score for each of the five assessments of the AISA battery was based on the ratings and answers obtained from each Soldier.

Table 37. Correlation between AISA KSAs and IPIP Constructs

	IPIP Scales				
	Extroversion	Agreeableness	Emotional Stability	Openness	Conscientiousness
Peer Leadership					
Interview Leadership	.05	-.03	.06	-.03	-.19
SBISE Leadership	-.06	.03	-.10	-.12	.02
Community Center Leadership	.17	.10	.21	-.04	-.14
DC Tour Leadership	-.08	-.02	-.05	-.12	-.11
Cultural Tolerance					
Interview Cultural Tolerance	.11	.13	.10	-.09	-.14
SBISE Cultural Tolerance	-.09	-.08	-.15	-.08	-.15
Concern for Soldier Quality of Life					
SBISE Concern for QoL	.10	.23*	.06	-.09	.06
Relate to and Support Peers					
Interview Relate to and Support Peers	.00	-.05	-.08	.02	-.17
Community Center Relate to and Support Peers	-.03	.04	.12	.01	-.16
DC Tour Relate to and Support Peers	.01	.09	-.19	-.20	-.21
Teamwork					
Interview Teamwork	-.11	.02	-.08	-.07	-.29**
Community Center Teamwork	-.13	-.09	.02	-.11	-.11
DC Tour Team work	.01	.08	.07	-.13	-.08

Note. ** indicates significance at the .01 level; * *S*ignificant at the 0.05 level
Underlined cells denote specific hypothesized relationships between the AISA KSA and the IPIP trait measure.

Correlations were run to evaluate the degree of relationship between the overall scores obtained from each test that is part of the AISA battery. Table 38 shows the correlation coefficients for each overall assessment score for AISA tests. Of interest are the significant positive relationships between overall score on the Semi-Structured Interview and Rater scores on the Community Center LGD exercise ($r = .25, p < .05$) and the DC Tour LGD exercise ($r = .28, p < .05$). These correlations indicate that individuals scoring well on the interview also tend to score well on the LGD exercises. Additionally, there was a significant relationship between the RBI and the rater scores on the Community Center LGD exercise ($r = .54, p < .01$). This relationship suggests that those scoring highly on the interpersonal skills measured by the RBI also frequently demonstrate the skills rated by the Community Center LGD exercise. The only remaining significant relationship identified between overall assessment scores is the relationship between the two LGD exercises which has been previously discussed in this report.

Table 38. Correlations of Overall Scores for AISA Battery

	Interview	SBISE	WCA	RBI	CC
Interview					
SBISE	.02				
WCA	.05	.11			
RBI	.25*	.19	.06		
Community Center LGD	.25*	.06	.06	.54**	
DC Tour LGD	.28*	.09	-.09	.13	.43**

* indicates correlation significant at the .05 level.
** indicates correlation significant at the .01 level.

Along with the correlation analyses, a multiple regression analysis was conducted to determine the amount of change in supervisor ratings that can be explained by a linear combination of the overall scores from each of the AISA assessments. Using the DC Tour LGD exercise, WCA, SBISE, RBI and Interview scores to predict average supervisor ratings of effectiveness shows no statistically significant predictive ability, $F(5,42) = .62$, n.s. An increase in predictive ability was observed when replacing the DC Tour LGD scores with the Community Center LGD scores, however results were still not statistically significant, $F(5,41) = 2.21$, n.s.

In addition to looking at the ability of the AISA battery to predict overall and mean ratings of effectiveness, we created composites from the supervisor rating dimensions based on previous research findings (Keenan, Russell, Le, Katkowski, & Knapp, 2005). Figure 16 shows how the composites were created (i.e., which rating dimensions constitute the composites). Composite scores were computed by simply averaging scores of the component rating dimensions. Table 39 contains the standardized beta coefficients for predicting supervisor overall and mean ratings of effectiveness along with the three ratings composites from overall test scores on the RBI, WCA, SBISE, Interview and Community Center LGD exercise. Results of the analysis indicate that the predictive ability for the AISA battery for the three composite scores are similar to those found in predicting overall and mean ratings of effectiveness. The combination of tests was found to predict a significant amount of variance in the Teamwork composite with an $R^2 = .22$. Table 40 also contains beta coefficients for predicting the ratings and composites using the DC Tour LGD exercise in place of the Community Center score. Using these assessments no significant predictive ability was found.

Figure 16. The Criterion Rating Composites

Rating Composite	Component Rating Dimensions
Teamwork	Supports Peers Exhibits Tolerance
Effort and Initiative	Effort Professionalism Professional Development Physical Fitness
Effort and Teamwork	Effort Professionalism Supports Peers Exhibits Tolerance

Table 39. Standardized Beta Coefficients for Predicting Ratings Composites from AISA Tests using Community Center LGD Scores

	Teamwork	Effort and Initiative	Effort and Teamwork	Overall Effectiveness	Average Effectiveness
RBI	-.27	-.03	-.17	-.28	-.16
WCA	-.03	-.17	-.15	-.28*	-.19
SBISE	-.15	.08	-.03	-.12	-.04
Interview	.42**	.27	.40**	.17	.44**
Community Center – LGD	.06	-.08	-.05	-.11	-.13
R^2	.22	.09	.16	.20	.21
$F_{(5,42)}$	2.28**	.84	1.53	2.07	2.21

* indicates significance at the .05 level; ** indicates significance at the .01 level.
Predictor variables entered in a single block.

Table 40. Standardized Beta Coefficients for Predicting Ratings Composites from AISA Tests Using DC Tour LGD Scores

	Teamwork	Effort & Initiative	Effort & Teamwork	Overall Effectiveness	Average Effectiveness
RBI	-.22	-.13	-.14	-.29	-.19
WCA	.06	-.11	.05	.10	-.05
SBISE	-.10	.12	.00	.02	.01
Interview	.12	.23	.16	.20	.23
DC Tour – LGD	.06	-.21	-.08	-.16	-.06
R^2	.07	.07	.04	.13	.07
$F_{(5,42)}$	.62	.64	.33	1.18	.62

* indicates significance at the .05 level; ** indicates significance at the .01 level.
Predictor variables entered in a single block.

Chapter 10: Conclusions and Recommendations

As the Army of today transforms into the Army of the future, interpersonal skills are becoming increasingly important. Unit focused stabilization will keep groups of people together for longer periods of time. This, coupled with an increased emphasis on small-team work, will engender an environment in which effective interaction between Soldiers is key. As such, the ability to identify and assess a Soldier's aptitude to work effectively with others is an important piece of future Army selection and assignment. The goal of the AISA battery is to measure the KSAs that are relevant to Soldiers' aptitude to work well with others as they carry out their mission. The Phase II SBIR effort was aimed at determining whether the innovative approach to measuring interpersonal skills was a valid method for selecting and assigning Soldiers to jobs that would require higher levels of interpersonal skills. From the findings outlined in this report, a number of conclusions can be drawn about the AISA. These conclusions, along with a set of recommendations for future activities to improve the battery, are the subject of this concluding chapter of the Phase II Final report.

The validation research for the AISA battery included a relatively small number of Soldiers and supervisor ratings (n = 95). Any assertions about the validity of the battery must be considered in the light of this limited sample. Because some of the predictors showed positive correlations with supervisor ratings, this suggests that the AISA as a whole as a concept for measuring interpersonal skills may hold promise, but requires further data collection and development to more firmly understand and establish the relationships observed.

A specific target for additional investigation is the SBISE. The SBISE shows significant positive relationships with mean supervisor ratings of effectiveness, suggesting that it may be a valid predictor of Soldier performance. However, the lack of relationship between the SBISE and other assessments in the AISA battery suggests that further investigation into the predictive relationship is required prior to employing the SBISE in a selection and assignment setting.

The WCA represents an attempt to measure a set of variables that are likely to become an increasingly significant element in interpersonal skills assessment. The increasing use of electronic communications will create the need for improved skills at using and interpreting email for all Soldiers. Given this, measures like the WCA will be important for use in the Army of the future. Unfortunately, more work is needed before the WCA in its current form can be applied to measuring a Soldier's aptitude to effectively interpret the interpersonal aspects of electronic mail. Future work should focus on clearly establishing the facets of electronic communication that are relevant to the interpretation of tone and intent. Additionally, efforts should be made to classify the types and frequency of electronic communication between Soldiers. Of promise is the small, positive relationship found between the WCA and supervisor estimates of communication ability. It is hoped that this relationship can be further solidified through additional validation data collection.

The ability of the AISA battery to predict a significant amount of variance in the Teamwork score composite is an important finding of the research effort. Of the 13 dimensions rated by supervisors, the Teamwork composite contains the two ratings that are most directly related to the type of interpersonal KSAs that are the target of the AISA. The Teamwork

composite includes supervisor ratings of the Soldier's aptitude to support peers and the degree to which the Soldier exhibits tolerance towards others. These two scales are directly related to personal characteristics that were identified in the Phase I effort as essential to effective interpersonal performance, and as such may serve as a valid surrogate for a full measure of interpersonal skill. It is reasonable to expect that the Supports Peers and Exhibits Tolerance ratings are less susceptible to ratings contamination introduced by non-interpersonal factors than other dimensions rated in the validation research. Many of the dimensions rated are relevant to the Soldier's generalized aptitude to perform effectively as a Soldier, but may not be related to their aptitude to function well in interpersonal situations. As such, ratings on these dimensions may not provide accurate estimations of the KSAs the AISA is intended to measure. However, the Teamwork composite attempts to remove some of the non-interpersonal aspects of performance. The ability of the AISA battery to account for a significant amount of score variance in this composite suggests that the battery may be a valid measure of a Soldier's interpersonal aptitude.

Finally, in Chapter 1 of this report we outlined our belief that both general mental ability (GMA) and trait dispositions have effects on knowledge and demonstration of interpersonal skills. The research team believes that trait dispositions may have a direct effect on an individual's skill in interpersonal situations and a residual effect on one's ability to perform in interpersonal situations. To test the effects of trait dispositions we included the IPIP measure of the Big Five personality traits and examined the relationship between our KSA measures and the personality variables measured by the IPIP. As detailed previously in this report, there appears to be no relationship between AISA measures of interpersonal skills and the personality factors measured by the IPIP marker. While our model (Figure 3) proposes a residual effect of trait disposition on performance level, it does not account for the impact of situational variables on both interpersonal skill and performance level. The validation sample used in this effort consisted primarily of Soldiers who had recently returned from a combat deployment. The intense emotional experiences associated with combat and the atypical skills that were utilized in that setting may have impacted the AISA measures of interpersonal skill by priming Soldiers to respond in a certain way that would be appropriate for a combat situation and overemphasize only a portion of their normal personality characteristics.

Recommendations

The research and development conducted in the Phase II SBIR yielded a great deal of insight into the measurement of interpersonal skill and produced a significant first step in measuring interpersonal KSAs for Army Soldiers. However, based on the results of the current validation effort, it is clear that additional work is needed for the battery to become a fully deployable selection and assignment tool. Additionally, this effort raised a number of issues and questions to be further investigated in the general area of interpersonal skills assessment and specifically interpersonal skills assessment using the AISA measures. The following paragraphs describe the research questions identified by the Phase II effort and provide our recommendations for future development of the AISA.

Of particular interest as an ongoing research question is the effect of recent combat experience on the interpersonal skills of Soldiers. The research team hypothesized that due to the

recent combat experiences of the Soldiers in the validation sample they may have been primed to exhibit a specific subset of responses to the predictor measures. Identifying and describing the specific influences (both short term and long term) that such extreme situations governed by rigid methods of interpersonal interactions have on the interpersonal skills of Soldiers would provide insight into the measurement of such skills.

In addition to the influence of combat experience on interpersonal skill, additional research is needed to understand the facets of written (specifically electronic) communication that are relevant to the interpretation of tone and intent. Anecdotal evidence suggests that individual differences exist in the ability to accurately interpret the tone and intent of an email message, but the WCA appears to have failed at measuring these differences. Research to better target the WCA on specific aspects of email that may be identifiable by those with higher levels of skill in understanding the tone and intent of email message would help in developing a more valid and reliable measure of the differences in this ability.

Another question identified in the current effort is associated with the construct of cultural tolerance. In the current research effort cultural tolerance proved to be an elusive construct to measure. As the frequency of inter-cultural interactions increases for Soldiers the ability to successfully navigate such situations will take on increased importance. Additional investigation is needed into the construct of cultural tolerance to identify better methods for measuring the ability of specific individuals to succeed in cross cultural encounters.

Finally, as discussed in Chapter 3 of this report, the supervisor performance rating scales used as the criterion measure in the validation study proved to be of low reliability and as such were not the ideal tool for validating the AISA measures. Future efforts at validating the AISA measures or approaches should seek to improve the reliability and validity of the criterion measures to eliminate this as a limitation on the measured validity of the predictor assessments.

Overall, the AISA represents a step in the right direction toward measuring Soldier interpersonal KSAs. The combination of methods provides a well-rounded and unique approach to measuring both interpersonal skill knowledge and the ability to implement that knowledge in specific situations. While additional research is certainly required before fully implementing the AISA for selection and assignment, it is clear from this research that the AISA and the lessons learned provide a roadmap for the assessment of a set of KSAs that will gain importance in the Army of the future.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695-716.

Bigelow, J.D. (1991). Giving an action exam: An evolving art. In J.D. Bigelow (Ed.), *Managerial skills: Explorations in practical knowledge.* Newbury Park, CA: Sage.

Bowden, T., Laux, L., Keenan, P., & Knapp, D. (2003). *Identifying and Assessing Interaction Knowledges, Skills and Attributes for Objective Force Soldiers* (Study Note 2004-01). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Brockson, S. (1999). *Selection and Teams: Effects of Selection Methods on Team Outcomes.* Unpublished master's thesis, University of Tennessee Chattanooga, Chattanooga.

Burlas, J. (2004, February 13). Force Stabilization means increased readiness. Army News Service. Retrieved November 4, 2005 at http://www.dcmilitary.com/army/ pentagram/9_06/national_news/27541-1.html

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of performance. In N. Schmitt, W.C. Borman, & associates (Eds.) *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass Publishers.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Sub-group differences in test performance and face validity. *Journal of Applied Psychology*, 82(1), 143-159.

Collins, J.M. (2000). *A meta-analysis of the construct validity of assessment centers.* Retrieved November 7, 2005 at http://www.cj.msu.edu/~faculty/collinsmeta.html.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, 54, 853-863.

Ferris, G.R., Witt, L.A., & Hochwarter, W.A. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology*, 86, 1075-1082.

Flesch, R. (1974). *The art of readable writing.* New York: Harper and Row.

Ford, L.A., Campbell, R.C., Campbell, J.P., Knapp, D.J., & Walker, C.B. (2000) *21st Century Soldiers and Noncommissioned Officers: Critical Predictors of Performance* (Technical Report 1102). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hogan Assessments (2005). *Hogan Assessments Glossary*. Located at
http://www.hoganassessments.com/Glossary.aspx. Downloaded August, 24, 2005.

HR Guide to the Internet (2000). *Personnel selection: Methods: Assessment centers*. Retrieved
November 7, 2005 at http://www.hr-guide.com/data/G318.htm

International Personality Item Pool (2001). *A Scientific Collaboratory for the Development of
Advanced Measures of Personality Traits and Other Individual Differences*
(http://ipip.ori.org/). Internet Web Site.

Keleman, K.S., Garcia, J.E., & Lovelace, K.J. (1990). *Management incidents*. Dubuque, IA:
Kendall/Hunt Publishing Co.

Keleman, K.S., Lovelace, K.J., & Garcia, J.E. (1991). Experiences with interaction skills
development: A situational approach. In J.D. Bigelow (Ed.), *Managerial skills*. Newbury
Park, CA: Sage Publishing Co.

Kilcullen, R.N., Mael, F.A., Goodwin, G.F., & Zazanis, M.M. (1999). Predicting U.S. Army
Special Forces field performance. *Journal of Human Performance in Extreme
Environments*, 4, 53-63.

Kilcullen, R.N., Putka, D. J.., McCloy, R.A. & Van Iddekinge, C. H. (2005). Development of the
Rational Biodata Inventory (RBI). In D.J. Knapp, C.E. Sager, & T.R. Tremble (Eds.)
*Development of experimental Army enlisted personnel selection and classification tests
and job performance criteria* (Technical Report 1168). Alexandria, VA: U.S. Army
Research Institute for the Behavioral and Social Sciences.

Klein, H.A., Pongonis, A., & Klein, G. (2000). *Cultural barriers to multinational C2 decision
making*. Presented at the 2000 Command and Control research and Technology
Symposium, Naval Postgraduate School, Monterrey, CA.

Knapp, D.J., Burnfield, J.L., Sager, C.E., Waugh, G.W., Campbell, J.P., Reeve, C.L., Campbell,
R.C., White, L.A., & Heffner, T.S. (2002). *Development of Predictor and Criterion
measures for the NCO21 research program* (Technical Report 1128). Alexandria, VA:
U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., McCloy, R. A., & Heffner, T. S. (2004). *Validation of measures designed to
maximize 21st century Army NCO Performance* (Technical Report 1145). Alexandria,
VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Sager, C. E., & Tremble, T. R. (Eds.). (2005) *Development of Experimental Army
Enlisted Personnel Selection and Classification Tests and Job Performance Criteria*
(Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the
Behavioral and Social Sciences.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants'
reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, 47,
715-738.

McCormick, N.B., & McCormick, J.W. (1992) Computer friends and foes: Content of undergraduates' electronic mail. *Computers in Human Behavior, 8,* 379-405.

McCrae, R. R., Costa, P. T., Terracciano, A.; Parker, W. D., Mills, C. J., De Fruyt, F., & Mervielde, I. (2002). Personality trait development from age 12 to age 18: Longitudinal, cross-sectional and cross-cultural analyses. *Journal of Personality and Social Psychology.* 83(6), Dec 2002, 1456-1468.

McCrae, R. R., Costa, P. T., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M., et al. (2000). Nature over nurture: Temperament, personality, and life span development. J*ournal of Personality and Social Psychology.* 78(1), 173-186.

Motowidlo, S.J., Dunnette, M. D., & Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 6,* 640-647.

Phillips, M.M. (2004). Simper nice: Before heading to Iraq, Marines learn people skills. *Wall Street Journal* (Eastern edition). Retrieved November 4, 2004.

Pruitt, D.G., & Rubin, J.Z. (1986). *Social conflict: Escalation, stalemate and settlement.* New York: Random House.

Russell, T.L., Crafts, J.L., & Brooks, J.E. (1995). *Intercultural communication requirements for Special Forces teams* (Research Report 1683). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Bemer, J. G., & Seaton, F. W. (1977). Job samples vs. paper-and-pencil trade and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187-197.

Smither, I. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology,* 46, 49-76.

Sproull, L. & Kiesler, S. (1991) *Connections: New ways of working in the networked organization.* Cambridge, MA: The MIT Press.

Sproull, L. & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science,* 32, 1492-1508.

Stokes, G.S. (2001). *Finding the past that is relevant for the future: Biodata form development and validation.* Presentation at the October 1999 meeting of the Personnel Testing Council of Metropolitan Washington, Washington, DC.

Witt, L.A., & Ferris, G.R. (2003). Social skill as a moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology*, 88, 809-820.

Appendix A: Definitions of Interpersonal KSAs

Relating to and Supporting Others

Ability to Relate to and Support Peers. The degree to which the individual treats peers in a courteous, respectful, and tactful manner. Provides help and assistance to others. Backs up and fills in for others when needed. Works effectively as a team member.

Amicability. The degree of pleasantness versus unpleasantness exhibited in interpersonal relations. Exhibits goodwill towards others and an absence of antagonism. Is tactful and helpful rather than defensive, touchy, and generally contrary.

Concern for Soldier Quality of Life. Is sensitive to others' priorities, interests, and values, and tries to assist them in making their personal and family life better.

Conflict Management

Conflict Management. The degree to which the individual encourages and supports different perspectives, avoids harmful conflict, constructively addresses disagreements that undermine team performance, and does not allow conflicts with others in ways that preserve good relations and enhance trust.

Cultural Tolerance

Cultural Tolerance. The degree to which an individual demonstrates tolerance and understanding of individuals from other cultural and social backgrounds, both in the context of the diversity of U.S. Army personnel and interactions with foreign nationals during deployments or when training for deployment.

Dependability

Dependability. The person's characteristic degree of conscientiousness. Is disciplined, well organized, planful, and respectful of laws and regulations.

Teamwork

Team Orientation. The degree to which an individual identifies with the team and other team members and works to boost team morale and increase the team bond by creating and maintaining a supportive work environment; willingness to put the needs of the team ahead of personal needs.

Coordination. The ability to work interdependently to reach task completion, share information and effort, and work together with others. Can adjust own time and work activities to ensure interdependent tasks are completed effectively.

Cooperativeness in Problem-Solving. The ability to take advantage of multiple perspectives to find effective solutions to problems.

Adaptability/Flexibility

Adaptability/Flexibility. The degree to which an individual is able to respond to rapidly changing situations (e.g., assignments, relocation, new Soldiers) and accept new roles.

Social Perceptiveness

Social Perceptiveness. The degree to which an individual is able to monitor own and other's emotions, discriminate among them, and use the information to guide one's thinking and actions, allowing one to work cooperatively with others. Is aware of how own behavior impacts others.

Communication Ability

Oral Communication Skills. The ability to speak clearly and precisely so that others can easily understand. The ability to adapt speaking style and comments to the audience, as appropriate and to listen effectively while focusing on the person communicating. The ability to incorporate appropriate non-verbal messages to clarify and enhance the message and to accurately interpret nonverbal signals of others.

Written Communication. The ability to write clearly so that message is understood by the reader. Is sensitive to the limitations of written communication (e.g., email) and carefully phrases message so that the intent can be clearly understood by the receiver.

Peer Leadership

Acts as a Role Model. Exhibits self-confidence and a positive attitude. Presents a positive and professional image of self and the Army even when off duty.

Helping Others. The ability to help other team members to improve performance. Willingness to provide assistance as needed and to guide and tutor others on technical matters.

Task Leadership. Ability to help keep the team focused on the team's assignment or mission, working with team members to react to changes and to ensure that conflicts do not hinder mission achievement.

Appendix B: AISA Software Description

        The AISA installation software places a shortcut on the computer desktop to launch the assessment software. The AISA computerized assessment battery is administered by double clicking the AISA icon on the computer's desktop. The AISA opens up a log in window (see Figure B1) which prompts the test taker to enter his or her User ID. The User ID is a six digit number unique to the participant. This identification number is used to designate the output files that capture the users answers stored in the User Answers directory. The User must enter the User ID number two times and then click "Log In" to take into the AISA battery.
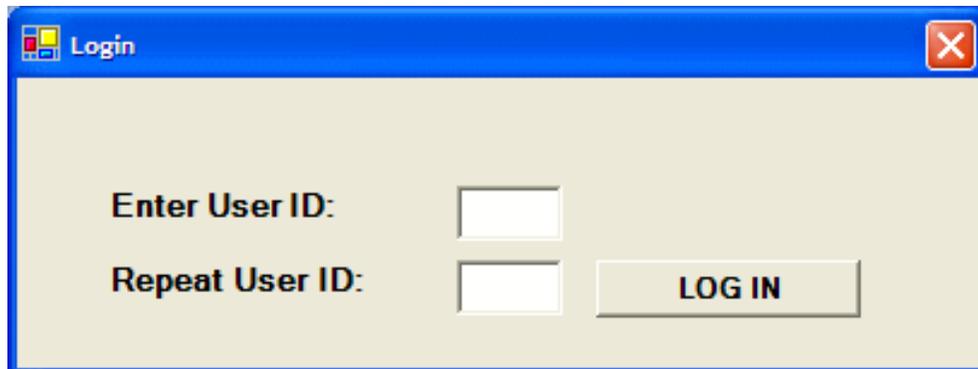
Figure B1. AISA log in screen.

        After logging in, the user is taken to an introduction screen which explains the purpose and importance of the assessment battery (see Figure B2). After the user reads and understands the introduction text, he or she clicks "Continue" to be taken to the assessment selection screen.
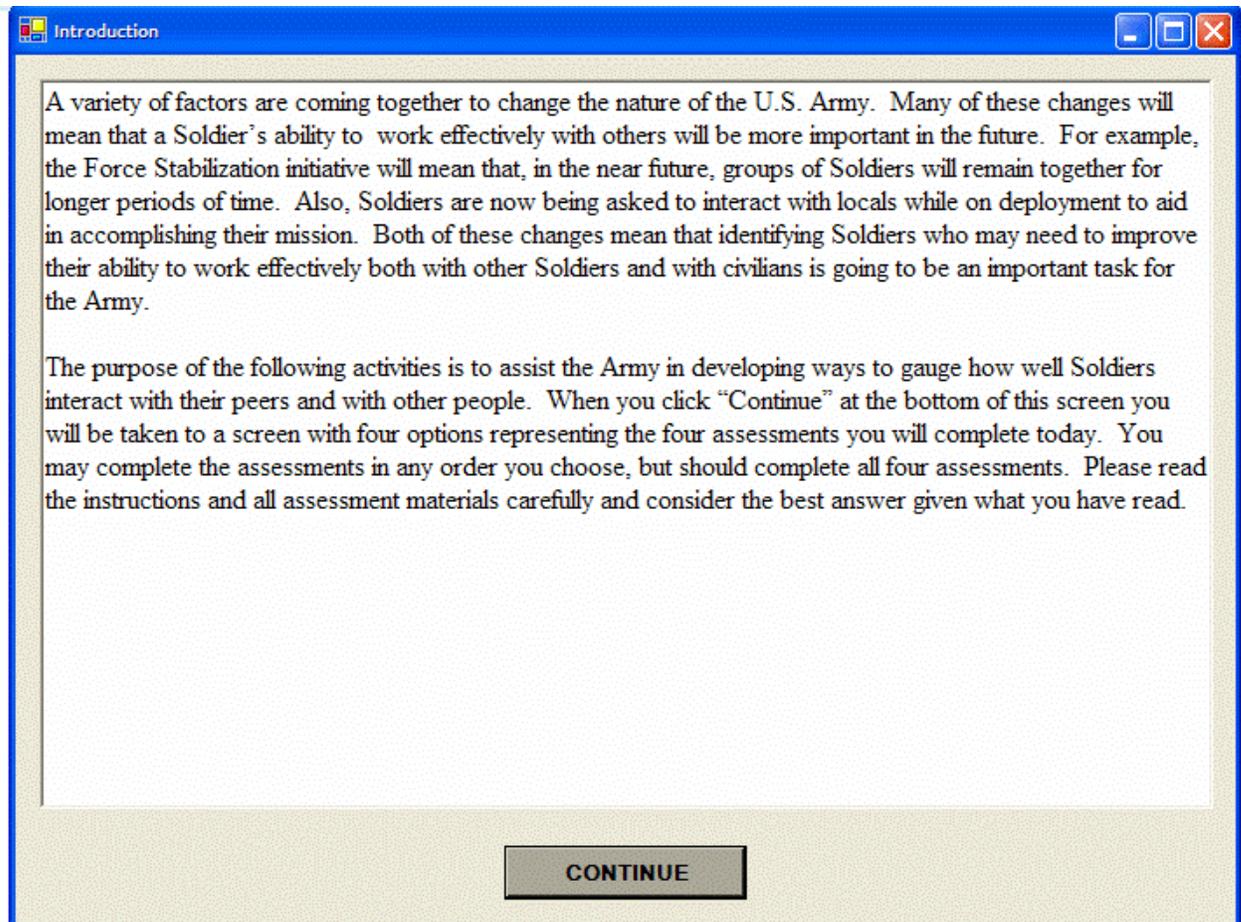
Figure B2. AISA introduction text.

The Assessment Selection screen (see Figure B3) depicts the assessments the user completes as part of the AISA battery. By clicking on a particular option the user enters that assessment and is taken to the specific instructions that correspond to the chosen assessment. The four computerized assessments that are administered as part of the AISA are the Rational Biodata Inventory (RBI), the Scenario Based Interpersonal Skills Evaluation (SBISE) and the Written Communications Assessment (WCA). Respondents complete the battery in 105-135 minutes. Specifically, the RBI takes 5-10 minutes, the SBISE takes 60-90 minutes, and the WCA takes 30-45 minutes.
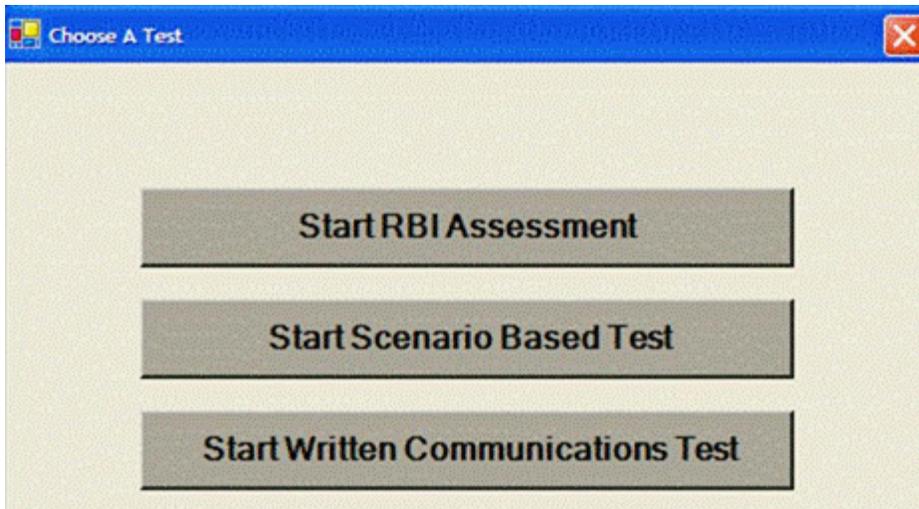
Figure B3. Assessment selection screen.

RBI Administration

To begin the Rational Biodata Inventory the user clicks on the top button on the assessment selection screen. When the user selects "RBI," the AISA software opens a new window that contained the RBI items. The AISA software presents the user with the 31 items in five item sets. The user selects the desired response from the drop down menu to the right of the assessment items and when the user selections are complete for a set of items, the user then clicks the "Submit Answers" button. It is possible for the user leave an RBI item incomplete, however, the system informs him or her that items are blank and asks if he or she wishes to continue submitting the answers or return and complete the unfinished items. When the test taker submits his or her answers to the final item set, the AISA software opens up the Assessment Selection window and allows users to choose the next assessment to complete.

Scenario Based Interpersonal Skills Evaluation (SBISE) Administration

When the participant selects the Scenario Based Interpersonal Skills Evaluation (SBISE) in the assessment selection screen, an instruction screen opens. After reading the instructions, the user clicks on the "Next" button to begin taking the Scenario Based assessment. There are two main screens to complete the Scenario Based assessment, the Video Interface (see Figure B4) and the Questions Interface (see Figure B5). When the assessment launches, the Video Interface opens and the first scenario animation begins. In the Video interface, the user can pause and stop the animation but cannot close the interface until he or she has viewed the animation. Once the scenario animation completes, the Video Interface closes and the Questions Interface opens to display the assessment items related to the previous video.

Figure B4. Sample of SBISE video player.

       The Scenario Based test contains two primary question types: Multiple-choice and Rating. For multiple-choice items, the user is shown a single item in the question box at the top of the interface with the possible answer options displayed in the lower box of the interface. The user selects his or her preferred option from a drop down list in the lower right hand side of the interface and clicks the "Submit Answers" button to move to the next question. If the user wants to replay the scenario animation, a button on the lower left of the questions interface reopens the video interface and replays the most recent animation.
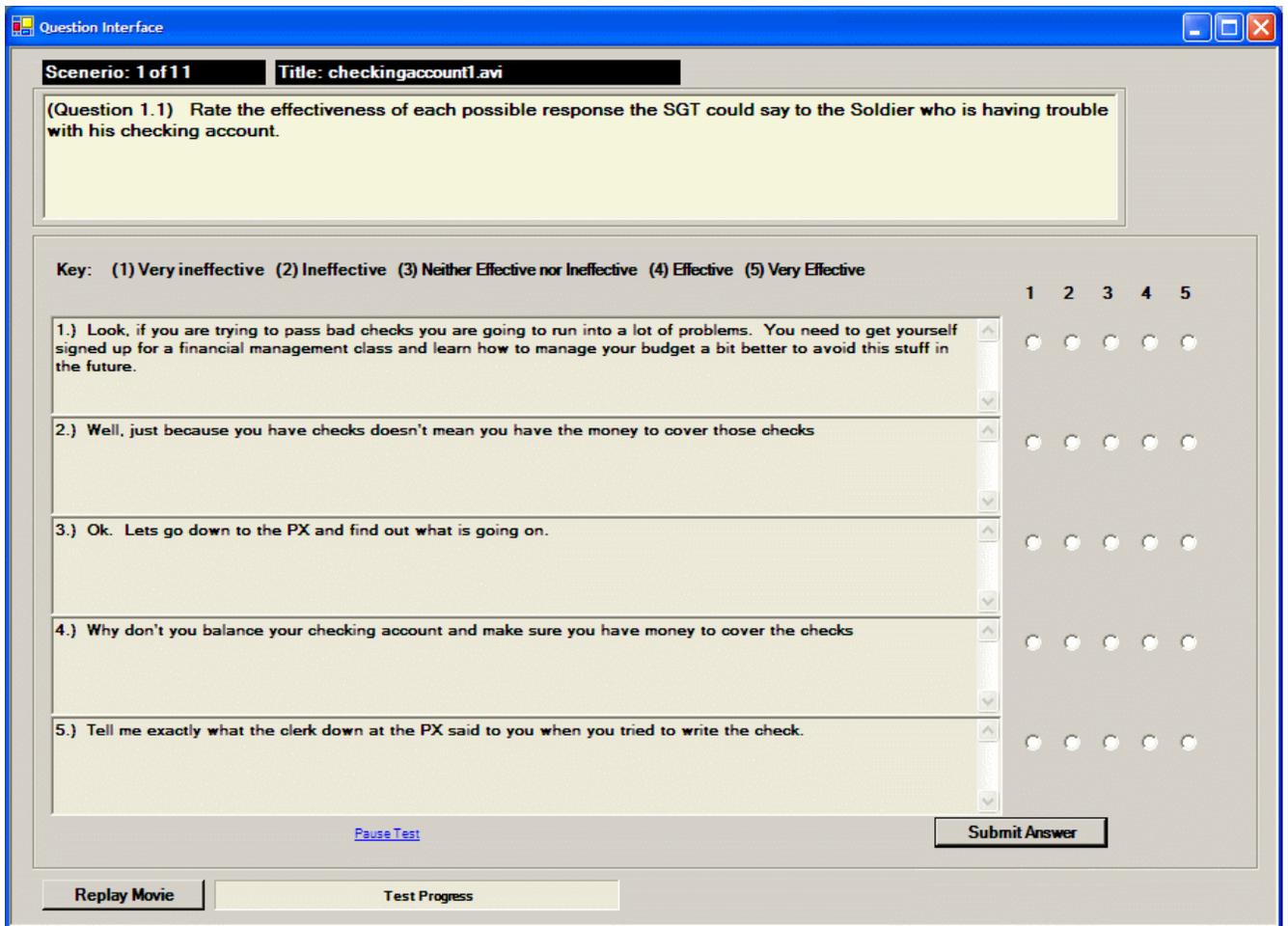
Figure B5. SBISE rating questions screen.

Written Communications Assessment Administration

        To take the WCA, the user selects the bottom button on the assessment selection screen. As with other assessments, when the selection is made the AISA software opens an instruction window with information telling the user how to complete the WCA. After reading the instructions, the user clicks the "Next" button to continue with the assessment. The user interface for completing the WCA is similar to those used to complete other assessments (Figure B6). In the WCA, the user is presented with a series of scenarios comprised of a set of emails that represent a set of communications about a given subject. The user is presented with the emails in the upper half of the user interface and is asked to read the emails and respond to a series of questions about the emails. After reading the emails, the user responds to assessment items by selecting the appropriate answer from the options shown. Once an option is selected, the user clicks the "Submit Answer" button to move to the next item.
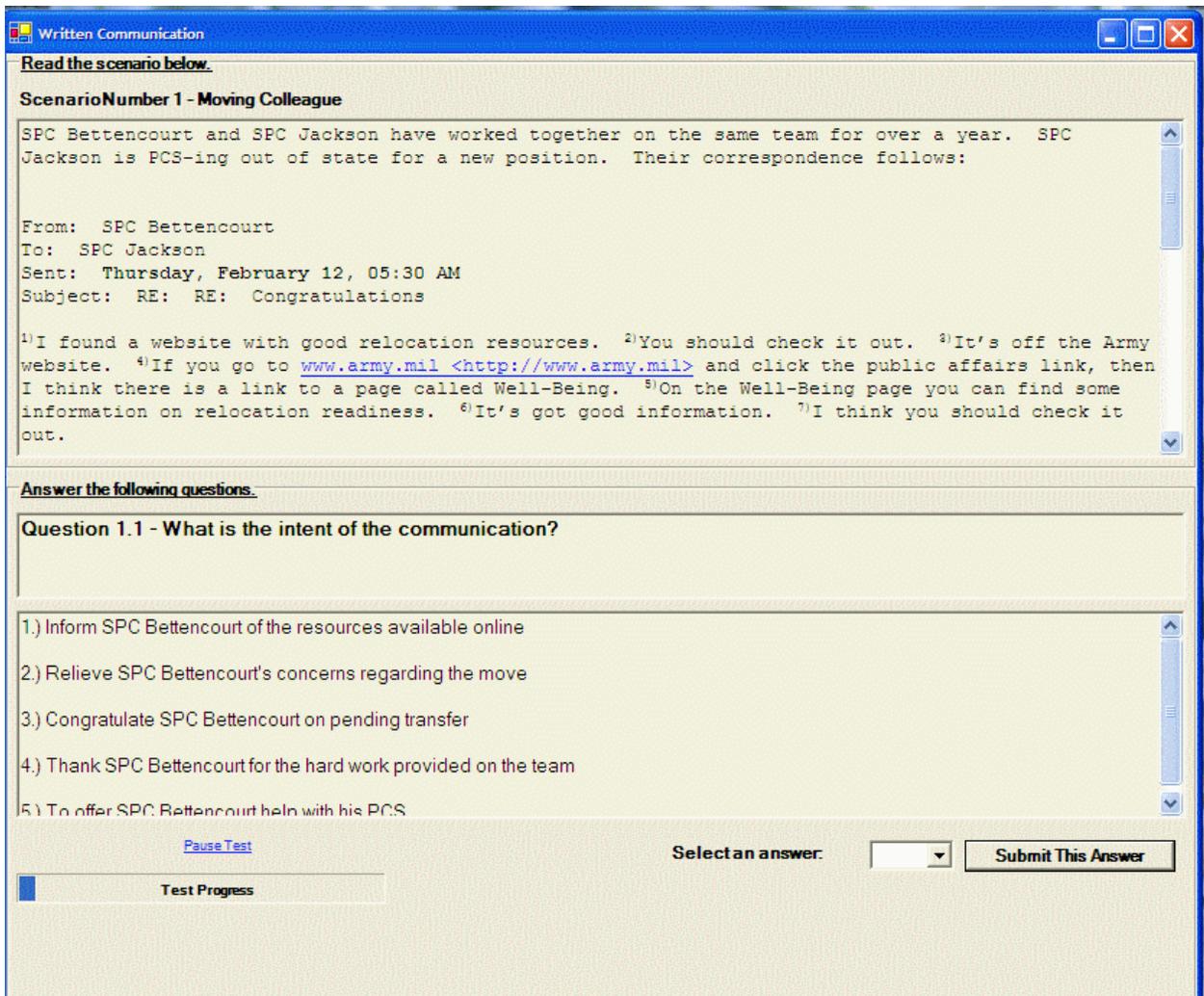
Figure B6. WCA interface.

User Instructions

Below is a list of important points that users are told to remember when responding to computerized assessment items.

1. Once the user clicks the "Submit Answer" button at the bottom of the user interface, the answer selected for that item can not be changed.

2. If the user chooses not to answer an assessment item, the software will confirm that no answer is being entered for the particular item.

3. Each question interface provides the user with the ability to pause the assessment with a text link in the lower portion of the user interface. The pause functionality stops the assessment timer and opens a blank window which should be closed to return to the assessment.

4. A test progress bar is provided in the lower left of each question screen to enable the user to track his or her progress through the assessment.

5. If the user clicks the X in the upper right hand corner of any question interfaces, the program will for confirmation of the request to exit the program. If the user chooses to exit the assessment software, he or she must re-log in to continue testing and restart any assessment that was not completed at the time the software was closed.

An additional significant feature of the AISA software is that it provides the assessment administrator with two testing modes that can be utilized based on the end use of the test outputs. Prior to test administration the test supervisor should access the administrator settings of the AISA software on the testing computer and select either Selection Mode or Development Mode. The mode the AISA software determines the output reports that will be provided along with the method for accessing those output reports. In Selection Mode, the AISA battery output reports are stored in the backend database for later review by the test administrator. These reports can be viewed either on a Soldier-by-Soldier basis or presented as a table showing the scores of all Soldiers in the given database. Selection reports can also be viewed for all tests at once or for a single test at a time. These reports contain overall assessment level scores for each test completed by the user. It is recommended that Selection reports be used in the context of a group of test takers to rank order test takers on their overall scores within a given assessment. This rank order list of examinees, along with individual scores from the Stage Two assessments can then be used to assist in selection and assignment decisions where increased levels of interpersonal skills may improve job performance. However, due to the limited size of the validation sample (as discussed in this report) AISA scores should not be the sole evaluation factor used in selecting or assigning individuals for a given assignment.

There are two differences between the Selection report and the Development reports as provided by the AISA software battery. First, whereas Selection reports are stored for later review by the test administrator and not displayed to the examinee, Development reports are provided to the individual immediately following completion of the final assessment in the battery. These reports can either be saved to a file or printed so that the user has a set of scores that can be used to identify interpersonal skill areas that may need further development. The second difference between the two report types supports the use of the Development reports as a tool for interpersonal skill improvement. The Development report not only provides overall assessment scores for each test completed, but also provides scores for each individual interpersonal KSA as measured by a given assessment. Additionally, a document defining each KSA measured is provided with the Development report. The individual KSA scores, in conjunction with the furnished KSA definitions enable the user to seek out targeted development activities that can improve the specific KSA deficiencies as identified by the AISA battery. In contrast to the KSA level detail provided in the Development report, the Selection report only provides overall scores for each assessment that was completed by the examinee.